UNIVERSITAT DE BARCELONA

Desarrollo y utilización de herramientas bioinformáticas en el estudio de datos de secuenciación masiva: Análisis genómicos en arácnidos

José Francisco Sánchez Herrero

Universitat de Barcelona

Facultat de Biologia

Departamento de Genética, Microbiología y Estadística

Desarrollo y utilización de herramientas bioinformáticas en el estudio de datos de secuenciación masiva: Análisis genómicos en arácnidos.

José Francisco Sánchez Herrero

Barcelona, Septiembre 2019

UNIVERSITAT DE BARCELONA

Desarrollo y utilización de herramientas bioinformáticas en el estudio de datos de secuenciación masiva: Análisis genómicos en arácnidos.

Memoria presentada por **José Francisco Sánchez Herrero**
para optar al Grado de Doctor en Genética (HDK0S)
por la Universidad de Barcelona

Departamento de Genética, Microbiología y Estadística

El autor de la tesis
José Francisco Sánchez Herrero

Tutor y codirector
Dr. Julio Rozas Liras

Codirector
Dr. Alejandro Sánchez-Gracia

Barcelona, Septiembre 2019

"George emprendió solemnemente la tarea de educarme. Desde mi punto de vista, lo más importante era que dedicábamos parte de nuestro tiempo a la historia natural, y George me enseñaba con cuidado y minuciosidad cómo había que observar y tomar nota de lo observado en un diario. Mi entusiasta pero desordenado interés por la naturaleza se centró, pues descubrí que anotando las cosas se aprendía y se recordaba mucho mejor. Las únicas mañanas en que llegaba puntualmente a mi lección eran las dedicadas a historia natural."

– Gerald Durrell, *Mi familia y otros animales* (1956).

A mi mujer e hijos,
por hacer que cada día brille el sol.

A mis padres y hermano,
por mostrarme lo bonita que es la
vida y siempre estar cuando hace falta.

A mis directores de tesis,
por iniciarme y guiarme en el
camino de la ciencia y la investigación.

Abstract

There is a vast amount of information indexed in genomic databases from multiple species of organisms but there is a bias against some taxonomic groups for their relevance at social, sanitary and economical level. The progress and the reduction of sequencing technologies have allowed the implementation of these techniques in non-model organisms but still, it is neither straightforward nor cheap to obtain high quality genomic resources. Spiders, main group of interest of this thesis, are non-model organisms underrepresented in genomic databases. The availability of a new genome would shed light into relevant biological traits such as the presence of venom, silk or the adaptation to terrestrial ecosystems and the chemosensory system.

The main objectives of this thesis are to develop bioinformatics methods and to generate genomic resources for non-model organisms, specifically, spiders. We have developed the tool DOMINO for the development of molecular markers in non-model organisms from next generation sequencing data. This tool allows identifying markers at different taxonomic ranges that could be employed directly, amplified using PCR in other related samples or for the generation of sequence capture strategies. We have validated the software using computer simulations and empirical data to adjust, configure and maximize its precision and sensitivity. Also, we have generated a graphical user interface to improve the usability of the software among those users with limited expertise in programming languages.

We have also developed a genomic assembly of a representative spider of the genus *Dysdera* by combining multiple sequencing technologies. Using several descriptive statistics we determined the quality and completeness of the assembly and conducted the structural and functional annotation by *ab initio* and evidence-based predictions. We obtained a genome with a N50 of 38 kb that could not be improved because the complexity of the genome, it includes a high proportion of repetitive regions, nevertheless the quality of genome in terms of gene completeness was fairly good. Globally we have generated a very useful genomic resource not only for conducting studies of specific biological or evolutionary characteristics in this genus but also for other arachnids or arthropods.

Índice general

Índice de Figuras

Índice de Tablas

Índice de Boxes

Glosario

ab initio Expresión latina que significa «desde el principio». En ciencias se dice que un cálculo es *ab initio* cuando sólo asume leyes básicas y bien establecidas, excluyendo parámetros externos o modelos simplificadores.

de novo Expresión latina que significa «de nuevo, desde el inicio». Se diferencia de *ab initio* ya que puede partir de modelos o parámetros determinados y no sólo de leyes básicas.

e.g. Por ejemplo (del latín *exempli gratia*).

et al. Y colaboradores (del latín *et alii*).

in silico Expresión que significa "hecho por computadora o via simulación computacional". La expresión, acuñada a partir de las frases *in vivo* e *in vitro* del latín, esta determinada por el material del que están hechos los semiconductores que permiten almacenar información en un ordenador, el silicio.

Anotar Proceso de determinar la localización de los genes y de sus regiones codificantes en un genoma y de establecer su posible función.

Artrópodo El término artrópodo (*von Siebold*, 1848) está formado a partir de las palabras griegas artro (articulación) y podos (pie).

Bead Perlas o bolas magnéticas que permanecen posicionadas en los pocillos de la placa de secuenciación para permitir el proceso de amplificación y secuenciación del ADN.

Biomimesis Ciencia que estudia la naturaleza como fuente de inspiración de nuevas tecnologías innovadoras para resolver problemas humanos mediante modelos de sistemas (mecánica), procesos (química) o elementos que imitan o se inspiran en ella.

BUSCO Acrónimo del inglés *"Benchmarking Universal Single Copy Orthologs"*. Consiste en un *software* y en un conjunto de datos generados para poder estandarizar el proceso de medir la calidad y continuidad de ensamblajes genómicos o de un conjunto de proteínas. La idea se basa en utilizar la presencia de aquellos genes ortólogos de copia única, compartidos por todos los organismos de un rango taxómico de interés, como medida indirecta de la calidad y continuidad de las secuencias que se analizan.

Contig Fragmento continuo de ADN obtenido mediante el ensamblaje de un grupo de reads de longitud inferior.

Ensamblaje Proceso de construcción de la secuencia genómica a partir de fragmentos secuenciados de ADN.

Estenofagia Estatregia alimentaria basada en la preferencia limitada por un número restringido de tipos de presas.

Hipógeo Ambiente subterráneo con escasez o ausencia total de luz, con humedad constante y con escasez de oxígeno.

Islas Macaronésicas Nombre colectivo de cinco archipiélagos del Atlántico Norte, más o menos cercanos al continente africano: Azores, Canarias, Cabo Verde, Madeira e Islas Salvajes.

Mapear Proceso de alinear secuencias cortas de ADN a una secuencia de referencia.

N50 Estadístico empleado en bioinformática como medida de la longitud media de un conjunto de secuencias nucleotídicas, con mayor peso dado a secuencias más largas. El valor N50 corresponde a la longitud por la cual el 50 % de todas las bases en las secuencias están en una secuencia de longitud superior a este valor.

Primer Cebador de ADN empleado durante la amplificación de un fragmento de ADN de interés.

Read Secuencia de ADN básica mínima resultante del proceso de secuenciación.

Scaffold Fragmento de ADN producto de la conexión de varios *contigs*. Cuando varias parejas de *reads*, producidos mediante librerías de secuenciación con conocidos tamaños de inserto (corto, "Paired-end" (PE) o largo, "Mate Pair" (MP)), se encuentran mapeando cada par en *contigs* distintos, se pueden conectar ambos *contigs*. Se deja una distancia aproximada, normalmente rellenada con *Ns*, determinada por la longitud del inserto de las librerías.

Ómica Neologismo que en biología molecular se emplea, como sufijo, para referirse al estudio del conjunto o de la totalidad de algo, como el conjunto de genes (genómica), proteínas (proteómica), transcritos de un organismo (transcriptómica), etc.

Abreviaciones

ADN Ácido desoxirribonucleico.

ANM Marcadores moleculares anónimos (del inglés *"Anonymous nuclear marker"*).

ARN Ácido ribonucleico.

ATAC-seq Secuenciación de zonas accesibles de la cromatima (del inglés *"Assay for transposable accessible chromatin sequencing"*).

CDS Regiones codificantes de proteínas en el genoma (del inglés "Coding sequence").

ChIP-seq Secuenciación de zonas compactadas de la cromatina (del inglés *"Chromatin mmunoprecipitation followed by sequencing"*).

CRT Terminación del Ciclo Reversible (del inglés *"Cyclic reversible termination"*).

ddNTPs Didesoxinucleótidos.

dNTPs Desoxinucleótidos.

EPIC Marcadores moleculares diseñados entre exones (del inglés *"Exon-Primed Intron Crossing"*).

FPU Beca predoctoral del MECD para la Formación de Profesorado Univesitario.

Gb Gigabases o Gbp (mil millones de pb) (del inglés *"Giga base pairs"*).

GUI Interfaz gráfica (del inglés *"Graphical user interface"*).

HGP Proyecto del Genoma Humano (del inglés *"Human Genome Project"*).

Hi-C Secuenciación masiva de captura de la conformación cromosómica (del inglés *"High-throughput sequencing Genome-wide chromosome conformation capture"*).

HTS Secuenciación masiva o de alto rendimiento (del inglés *"High throughput sequencing"*).

kb Kilobases o kbp (mil pb) (del inglés *"Kilo base pairs"*).

Mb Megabases o Mbp (un millón de pb) (del inglés *"Mega base pairs"*).

MECD Ministerio de Educación, Cultura y Deporte del gobierno de España.

methyl-seq Secuenciación del estado de la metilación del ADN (del inglés *"DNA methylation sequencing"*).

MINECO Ministerio de Economia, Industria y Competitividad del gobierno de España.

MP Fragmentos de secuenciación pareados con un tamaño de inserto largo (del inglés *"Mate pair-end reads"*).

MSA Alineamiento multiple de secuencias (del inglés *"Multiple Sequence Alignment"*).

My Millones de años (del inglés *"Million years"*).

NCBI Centro Nacional de Estados Unidos de análisis biotecnológico (del inglés *"National Center for Biotechnology Information"*).

NGS Nuevas tecnologías de secuenciación (del inglés *"Next Generation Sequencing"*).

NPCL Marcadores moleculares en zonas codificantes de proteínas (del inglés *"Nuclear protein coding loci"*).

NRT Nucleótidos Terminadores Reversibles (del inglés *"Nucleotide Reversible Terminators"*).

ONT Empresa de secuenciación de *reads* largos *"Oxford Nanopore Technologies"*.

PacBio Empresa de secuenciación de *reads* largos *"Pacific Biosciences"*.

pb Pares de bases (del inglés *"base pairs"*).

PCR Reacción en cadena de la polimerasa (del inglés *"Polymerase chain reaction"*).

PE Fragmentos de secuenciación pareados con un tamaño de inserto corto (del inglés *"Pair-end reads"*).

RADseq Secuenciación masiva asociada a sitios de corte de enzimas de restricción (del inglés *"Restriction-site-associated DNA sequencing"*).

RNAseq Secuenciación masiva de transcritos de ARN (del inglés *"RNA sequencing"*).

rNTP Ribonucleótidos.

RRL Librerías reducidas representativas (del inglés *"Reduced representation library"*).

SBL Secuenciación por ligación (del inglés *"Sequencing by ligation"*).

SBS Secuenciación por síntesis (del inglés *"Sequencing by synthesis"*).

scSeq Secuenciación masiva del genoma de células individuales (del inglés *"Single-cell sequencing"*.

SE Fragmentos sencillos de secuenciación (del inglés *"Single-end reads"*).

SMRT Lectura en tiempo real de secuencia de una única molécula (del inglés *"Single Molecule Real Time"*).

SMS Secuenciación de molécula única (del inglés *"Single Molecule Sequencing"*).

SNA Adición de un nucleótido (del inglés *Single Nucleotide Addition"*).

UCE Marcadores moleculares ultraconservados (del inglés *"Ultraconserved elements"*).

UTR Regiones no codificantes de un exon (del inglés *"Untranslated Regions"*).

WGS Secuención de genomas completos (del inglés *"Whole genome sequencing"*).

Capítulo 1

Introducción

El ADN determina en primera instancia la información hereditaria y bioquímica de la vida tal y como la conocemos. La habilidad de poder caracterizar y medir su naturaleza y su variación es un imperativo de la investigación en biología [1, 2].

A lo largo de esta introducción recorreremos la historia de la secuenciación del ADN y destacaremos las aplicaciones de éstas metodologías en el mundo de la biología haciendo hincapié en la secuenciación de genomas para su estudio y caracterización o en la aplicación de estas metodologías en estudios evolutivos. Pondremos en contexto el trabajo que aquí concierne y destacaremos la importancia tanto de las tecnologías de secuenciación como de la bioinformática en dicho estudio, especialmente, de organismos no modelo.

1.1. Obtención de la secuencia genómica

La cronología de la secuenciación del ADN es una historia ocurrida en apenas unas décadas y repleta de puntos de inflexión y cambios de paradigma (Figura 1; Tabla S1). Para esta introducción, nos centraremos en aquellos puntos más importantes de las metodologías que se han empleado durante el desarrollo de esta tesis. Para una revisión más exhaustiva de la historia de la secuenciación se pueden revisar otros trabajos científicos [1, 3–6].

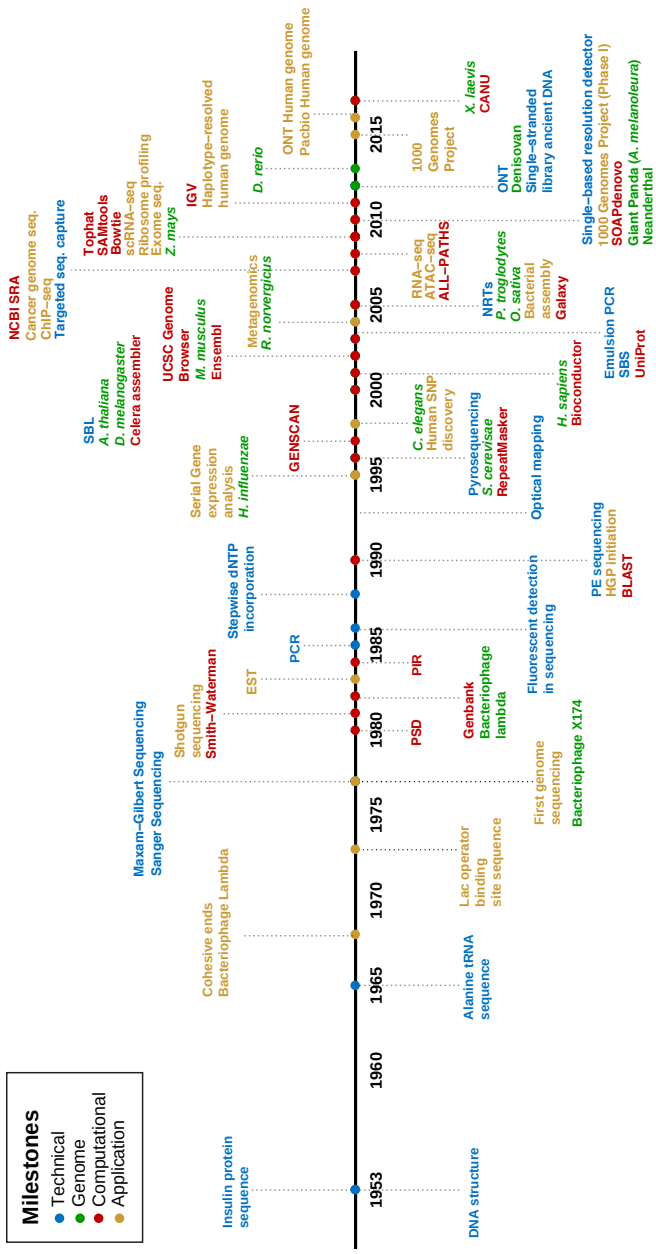


Figura 1: **Principales hitos en la historia de la secuenciación**. Fuente [4, 5]. Ver detalle en tabla S1.

Desde 1977, las tecnologías de secuenciación del ADN han evolucionado de forma muy dinámica y en la actualidad el panorama sigue en un continuo cambio. El lanzamiento de las primeras máquinas de secuenciación masiva a mediados de los 2000s y su posterior desarrollo propició que entre 2007-2012 el coste total de secuenciar una base se redujera en más de cuatro órdenes de magnitud [7] (Figura 2; Tabla S2). Estaba comenzando una nueva revolución en el mundo de la secuenciación de ADN, las denominadas tecnologías *"ómicas"*. Se puso de moda un nuevo término: secuenciación masiva, del inglés *"High Throughput Sequencing"* (HTS) o Secuenciación de nueva generación, *"Next generation sequencing"* (NGS) (Figura 2). Se incrementó el rendimiento de las nuevas tecnologías y se incorporaron numerosas innovaciones para afrontar las complejidades de los diferentes genomas y poder abordar nuevas cuestiones biológicas. Esto se traduce hoy en día en toda una serie de innovaciones tanto a nivel de investigación aplicada y clínica como de investigación básica [5]. Actualmente, las tecnologías permiten obtener la secuencia de un único fragmento de ADN de varias kilobases (kb), secuenciar un genoma humano a un coste tan reducido como 1000$ o generar información del genoma de una única célula.

1.1.1. Inicios de la secuenciación de ADN.

Desde 1953 se conoce la estructura del ADN gracias a los trabajos de Watson y Crick [11], a partir de datos cristalográficos de Rosalind Franklin y Maurice Wilkins [12]. Pero no fue hasta mucho más tarde que se tuvo la capacidad de "leer" este ADN. Las metodologías existentes para identificar secuencias proteicas (*e.g.* Insulina, Fred Sanger, 1953 [13]) no se podían aplicar al ADN. Era necesario el desarrollo de nuevas metodologías.

Tras unos primeros intentos poco eficientes [14, 15], no fue hasta 1977 que se desarrollaron métodos para obtener la secuencia de moléculas de ADN. Éstas permitían secuenciar cientos de bases en unas horas revolucionando el campo de la investigación en biología y popularizandose entre la comunidad científica muy rápidamente. Es lo que hoy en día conocemos secuenciación de primera generación (del inglés *"First generation DNA sequencing"*).

Los métodos, desarrollados por Sanger & Coulson [16, 17], se conocían como la técnica de terminación de cadena basada en didesoxinucleótidos (ddNTPs) (Box 1) (en inglés *"dideoxy technique"*) y el de Maxam & Gilbert [18] (en

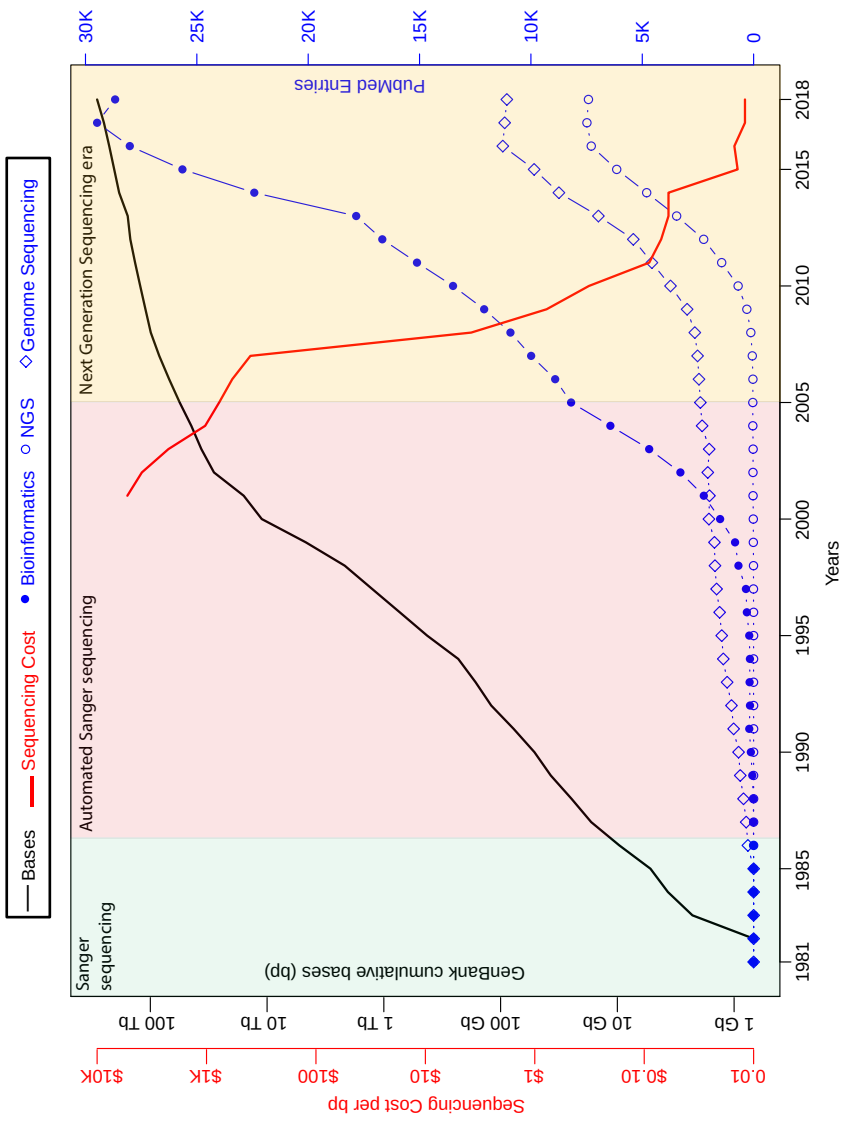
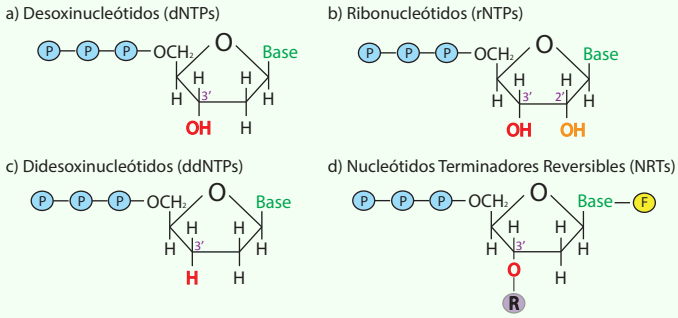


Figura 2: **Cronología de la historia de la secuenciación**. En negro, la cantidad de bases indexadas en la base de datos *GenBank* (1982-2019) en escala logarítmica [8, 9]; en azul, artículos científicos indexados en *PubMed* (1982-2019) [10] que contienen las palabras: *"Bioinformatics"*, *"Next generation sequencing"* y *"Genome Sequencing"*; en rojo, coste de secuenciación por par de bases (pb) (2001-2019) en escala logarítmica [7]. Ver detalles en tabla S2.

inglés *"chemical cleavage technique"*). Aunque difieren en la metodología de identificación de los nucleótidos, ambos se basan en la resolución por tamaño proporcionada por los geles de poliacrilamida que permite la separación de fragmentos que difieren por tan solo un único nucleótido. La precisión y robustez así como la facilidad de uso y sencillez de la técnica de Sanger la convirtieron en la tecnología más empleada en la secuenciación de ADN durante bastantes años siendo aún utilizada hoy en día.

Box 1: Ácidos nucleicos: nucleótidos originales y modificados



Los diferentes nucleótidos, monómeros de las moléculas de ADN (a) o ARN (b), están conformados por una pentosa, ácido fosfórico [P] (bolas azules) y una base nitrogenada, que se indica como *"Base"* (verde) y que pueden ser purinas (Adenina [A] o Guanina [G]) o pirimidinas (Citosina [C] o Timina [T] / Uracilo [U]). Los diferentes grupos hidroxilos que contienen la pentosa permiten la polimerización de la molécula y la clasificación de estos nucleótidos. Existen los desoxinucleótidos trifosfato (dNTPs) (a) que son los monómeros de ADN y que presentan un grupo hidroxilo (OH) en el carbono 3 (destacado en rojo) que permitirá la unión del siguiente mononucleótido. Por otra parte, los ribonucleótidos (rNTPs) (b), además del grupo 3'-OH contienen un grupo 2'-OH (marcado en naranja) que conferirá unas características distintivas a esta molécula de ARN. En algunas de las metodologías de secuenciación se han empleado análogos químicos de estos ácidos nucleicos. Por ejemplo los didesoxinucleótidos trifosfato (ddNTPs) (c), son análogos químicos de los dNTPs (a) que carecen del grupo 3'-OH, interrumpiendo la elongación de la cadena cuando se incorporan durante la síntesis de ADN. También encontramos los nucleótidos terminadores reversibles (NRTs) (d) que evitan la prolongación de la molécula al contener un grupo [R] (violeta) bloqueando el grupo 3'-OH. Además este método permite su identificación al constar de un fluorocromo (F) (amarillo) específico para cada base nitrogenada.

Numerosos avances facilitaron el uso de la secuenciación de ADN. En 1987, Smith, Hood y *Applied Biosystems* [19, 20] desarrollaron máquinas automatizadas de secuenciación Sanger que permitían una identificación mediante fluorescencia que generaba secuencias contiguas de fragmentos de hasta 1 kb (Figura 1). El desarrollo paralelo de técnicas de amplificación por reacción en cadena de la polimerasa (PCR, del inglés *"Polymerase chain reaction"*) [21, 22] y la de tecnologías de recombinación de ADN [23, 24] contribuyeron al desarrollo de la revolución genómica al permitir amplificar con mayor facilidad una cantidad suficiente de ADN necesaria en el proceso de secuenciación. Para poder determinar la secuencia genómica de organismos con tamaños genómicos grandes los investigadores tenían que utilizar técnicas como la denominada *"shotgun sequencing"* [25] mediante la cual fragmentos solapados de ADN eran clonados y secuenciados por separado. Fue también importante el desarrollo de la bioinformática (Box 2) que permitiría ensamblar *in silico* [26] las secuencias clonadas en un fragmento único denominado *contig*. Otros avances relevantes en el proceso de secuenciación fueron aquellos relacionados con las ADN polimerasas. Originalmente se utilizaba, por su habilidad para incorporar ddNTPs (Box 1), el fragmento *Klenow* de la ADN polimerasa, originaria de *Escherichia coli* [27]. Sin embargo, la aparición de técnicas de manipulación genéticas permitieron obtener nuevas polimerasas que incorporasen de forma más eficiente o con menor error los nucleótidos en las diferentes técnicas de secuenciación [28].

1.1.2. Secuenciación de nueva generación: *"Next generation sequencing"*

Desde las décadas de 1980s-1990s, muchos grupos de investigación exploraban alternativas a la secuenciación basada en electroforesis pero no fue hasta después del Proyecto del Genoma Humano (HGP) (del inglés *"Human Genome Project"*) (Box 3) que se desarrolló la secuenciación masiva o de nueva generación y se acuñó el término HTS y NGS, respectivamente. Las primeras plataformas de secuenciación masiva aparecieron en 2005 (*e.g. Roche 454*) [29] y muy pronto se incrementó en varios órdenes de magnitud el rendimiento de las diferentes tecnologías de secuenciación (Figuras 2-3; Tabla S3). En apenas unos años los laboratorios de tamaño moderado tuvieron acceso a este tipo de tecnologías. Se popularizaron en muchos ámbitos de la investigación y se dió paso a la era *"ómica"* de la biología.

Box 2: Breve historia de la Bioinformática

La historia del desarrollo de la tecnología de la secuenciación del ADN no se puede entender sin el progreso paralelo del desarrollo de herramientas bioinformáticas. Se precisó de numerosas innovaciones tanto a nivel de *hardware* como de *software* para permitir el procesamiento de la gran cantidad de datos y su análisis [30–32].

Como pionera de la bioinformática destaca Margaret Oakley Dayhoff (1925-1983) [2, 33], química-física de formación, que pronto descubrió el potencial de la computación aplicado al mundo de la química y el almacenamiento de secuencias. Inicialmente desarrolló un sistema para simplificar y estandarizar la nomenclatura de los aminoácidos (1968) [34] que aún se sigue utilizando hoy en día. Dayhoff también desarrolló modelos probabilísticos de substitución de aminoácidos, las matrices PAM (del inglés *"Point Accepted Mutation"* ó *"Percent Accepted Mutation"*) y creó en 1965 la primera base de datos de proteínas, el *"Atlas of Protein Sequence and Structure"* para almacenar las secuencias proteicas generadas.

Inicialmente, Emile Zuckerkandl y Linus Pauling (1962) observaron que los cambios de aminoácidos entre secuencias homólogas, en esta caso hemoglobina de diferentes vertebrados, eran debidos a la divergencia a partir de un ancestro común. Correlacionaron esta divergencia con el tiempo de divergencia de las especies obtenido a partir del registro fósil [2] y acuñaron el término *reloj molecular"* al afirmar que la tasa de cambio evolutivo era constante, aproximadamente, a lo largo del tiempo. Esta observación fue posteriormente explicada a nivel teórico por Kimura (1968) y es conocida como la teoría neutralista de la evolución molecular. Simultáneamente, Walter M. Fitch fue uno de los pioneros de la bioinformática en el campo de la biología. Empleó la bioinformática para la resolución de un problema biológico, en este caso, una reconstrucción filogenética, a partir de secuencias de citocromo C (1967). Además, definió el concepto de ortólogo, como secuencias homólogas resultante de un evento de especiación, y el concepto de parólogo, en este caso a partir de un evento de duplicación.

Pero para poder comparar secuencias era necesario solventar varios problemas, tanto conceptuales como computacionales. Fue necesario asignar un *"valor evolutivo"* a este tipo de cambios de aminoácido generados para poder así obtener el mejor resultado de una comparación de secuencias. Partiendo de esta base y del desarrollo de las matrices PAM, Needleman y Wunsch (1970) [35] desarrollaron el primer algoritmo de programación dinámica para alineamiento de proteínas y años más tarde, Da-Fei Feng y Russell F. Doolitle (1987) [36], desarrollaron la primera metodología de alineamiento múltiple conocida como alineamiento progresivo de secuencias. También se desarrolló el estudio de las reconstrucciones filogenéticas mediante programas basados en máxima verosimilitud (Felsenstein, 1981) [2].

Con el establecimiento de la idea, hoy conocida como dogma central de la biología, mediante la cual la información fluye del ADN al ARN y éste se traduce en proteínas, se produjo un cambio de paradigma que popularizó los estudios de ADN. Las técnicas de secuenciación del ADN comenzaron a desarrollarse y popularizarse, y pronto fue

necesario el desarrollo de *software* para el tratamiento de esta información y el desarrollo de diferentes bases de datos para almacenar la cantidad de información generada. Se comprendió la necesidad de fusionar, estandarizar y centralizar la información para facilitar el acceso y el intercambio entre la comunidad científica. Las bases de datos de secuencias almacenadas en *"European Molecular Biology Laboratory"* (EMBL), *"GenBank"* y *"DNA Data Bank of Japan"* (DDBJ) se unieron en 1986-1987. En la actualidad, esta unión aún se mantiene en lo que se denomina *"International Nucleotide Sequence Database Collaboration"* (INSDC) [37]. Tras el fallecimiento de la doctora Dayhoff en 1983, la base de datos *"Atlas of Protein Sequence and Structure"* se convertiría en la *"Protein Sequence Database"* (PSD) (1984) [33]. Años más tarde se fusionó con *"Protein Identification Resource"* (PIR) y a partir del año 2002, PIR-PSD se asoció con *"European Bioinformatics Institute"* (EIB) y *"Swiss Institute of Bioinformatics"* (SIB). Dieron origen a una única base de datos de secuencia y función de proteínas, conocida en la actualidad como *UniProt* [38, 39].

La gran difusión de las diferentes bases de datos no hubiera sido posible sin la llegada de Internet, que facilitó el acceso a los usuarios, así como el desarrollo del *software* necesario para el manejo y el análisis de grandes cantidades de datos [40]. Con el paso de los años y los numerosos avances técnicos, las secuencias generadas aumentaron exponencialmente (Figura 2), aproximándose a la *Ley de Moore* [5] y motivando la creación de repositorios centrales y de herramientas (*e.g.* BLAST [41]) para tener un acceso rápido y eficiente a esta información.

Más allá del manejo de la información primaria almacenada en las secuencias de los ácidos nucleicos, la bioinformática también se ha utilizado, sobre todo en el campo de la química, para obtener predicciones de la estructura tridimensional de moléculas y realizar simulaciones de dinámica molecular [2]. No entraremos en detalle en este ámbito de la bioinformática al quedar fuera del contexto de estudio de esta tesis doctoral.

La bioinformática es de naturaleza interdisciplinar y busca entender los procesos biológicos usando la capacidad de cálculo y computación de los ordenadores para la adquisición, manejo y análisis de la información biológica. Supone por tanto una intersección de la biología molecular, la biología computacional, las bases de datos informáticas, internet y el análisis de secuencias [42]. En los últimos años se ha hecho imprescindible su uso, sobre todo desde la disponibilidad de datos masivos (Figuras 1-2) y se han desarrollado numerosos *softwares* para poder trabajar con datos biológicos.

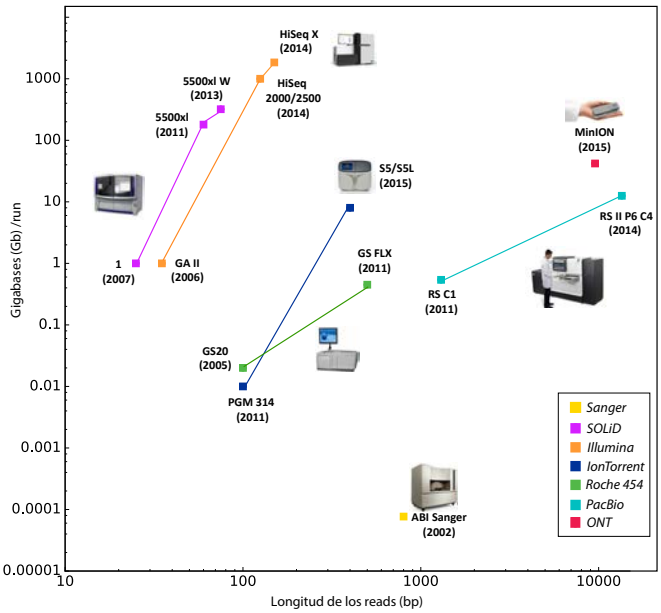


Figura 3: **Rendimiento de las principales metodologías de secuenciación**. En el eje Y, cantidad de bases secuenciadas en cada experimento, en escala logarítmica. En el eje X, la longitud media de las secuencias producidos por cada metodología, en escala logarítmica. Imagen adaptada de [43, 44]. Ver detalle en tabla S3.

A continuación exploraremos brevemente las metodologías más relevantes de secuenciación masiva, centrándonos en aquellas empleadas en esta tesis. Para una revisión en profundidad se pueden ver otros trabajos científicos [1, 3–6].

La secuenciación de nueva generación se diferencia, en grandes rasgos, de la secuenciación Sanger por su capacidad de paralelizar y realizar múltiples reacciones a la vez. En vez de un tubo de ensayo por reacción, los fragmentos de ADN son inmovilizados en una superficie y los reactivos para la reacción son expuestos (Box 4). La amplificación del ADN fijado permite que la señal emitida pueda ser diferenciada del ruido de fondo [4]. Por otra parte, el incorporar millones de centros de reacción, donde se realiza la secuenciación de un fragmento de forma individual, a lo largo de la superficie de fijación permite realizar reacciones de amplificación simultáneas. Para cada molécula fijada, se producirá de forma paralela la obtención de la secuencia de ADN.

Box 3: Proyecto Genoma Humano (HGP)

No se puede entender la historia de la secuenciación de ADN sin conocer la relevancia que tuvo el proyecto de secuenciación del genoma humano (HGP) [45, 46]. Fue un proyecto muy ambicioso, iniciado en los años 1990 y financiado por numerosas agencias públicas ($3000 millones en total), con múltiples colaboraciones internacionales. Este proyecto generó muchas dudas y escepticismo ya que se dudaba si los beneficios obtenidos con este proyecto iban a ser mayores que los costes. En la actualidad, no solo se han obtenido grandes beneficios tanto médicos como de investigación básica, sino que ha sido el punto de inflexión de múltiples tecnologías y en la forma de estudiar y abordar tanto enfermedades como otras cuestiones biológicas. No entraremos en detalle en los avances y beneficios del HGP en esta tesis, aunque son múltiples [5, 46, 47], sino que nos centraremos en los avances y el impacto que tuvo en la historia de la secuenciación del ADN y de la bioinformática.

La secuenciación del HGP se realizó íntegramente utilizando una secuenciación escalonada basada en *"shotgun sequencing"* [25] y denominada *"hierarchical shotgun sequencing"* [45]. Se generaron clones del genoma que se secuenciaron individualmente con metodología Sanger en máquinas automatizadas. La necesidad de generar múltiples pasos encarecía y ralentizaba el proceso. Se realizaron múltiples avances con el objetivo de incrementar la eficiencia del proceso. Se produjo un cambio de los *primers* marcados a terminadores marcados diferencialmente que permitió el poder hacer toda la reacción en un único pocillo [48]. Fue necesario por tanto una nueva ADN polimerasa, modificada del *bacteriófago T7*, que incorporaba estos nucleótidos de forma mucho más eficiente [49]. Se optimizaron procesos como la amplificación, que se volvió linear para así facilitar la miniaturización [50]; la purificación del ADN [51] o la lectura e interpretación de resultados mediante electroforesis capilar [52], que eliminó la necesidad de utilizar geles de acrilamida.

Durante el desarrollo del HGP surgió una empresa, *Celera Genomics*, con una metodología y aproximación completamente diferentes que generó una gran competencia en ser el primero en secuenciar el genoma humano. Fundada por Craig Venter, la empresa desarrolló la metodología de *"Whole Genome Shotgun Sequencing"*. Mediante esta tecnología y a diferencia de la anterior, se fragmenta el ADN en múltiples fragmentos que se secuencian de forma individual y luego mediante la bioinformática se pueden ensamblar de forma inequívoca. Fue necesario una capacidad de computación y manejo de los datos que era completamente desconocida hasta la fecha. Grandes empresas tecnológicas trabajaron conjuntamente para el desarrollo de este tipo de ordenadores y en el desarrollo de *software* bioinformático. Testada inicialmente con el genoma de *Drosophila melanogaster*, esta técnica pasó a emplearse en la secuenciación del genoma humano. Utilizando la información pública previa, en tan solo 9 meses se consiguió secuenciar y ensamblar el genoma humano mediante este proyecto de financiación privada [53].

Es indudable por tanto el papel catalizador que tuvo este proyecto de secuenciación del genoma humano, mediante capital público o privado, tanto en el avance de las tecnologías de secuenciación como en el desarrollo de la bioinformática.

Otra de las diferencias principales respecto a la secuenciación Sanger es que la determinación de la secuencia se realiza en cada ciclo de reacción mediante técnicas distintas:

- Por emisión de un fluoróforo, dependiente de ADN ligasas, conocida como secuenciación por ligación [54] (SBL, del inglés *"Sequencing by ligation"*) y empleada por plataformas como *SOLiD* o *Complete Genomics*;
- Por cambios en la concentración de los reactivos, dependiente de ADN polimerasas, conocida como secuenciación por síntesis (SBS, del inglés *"Sequencing by synthesis"*).

Puesto que son las más comunes en el mercado y las empleadas en esta tesis, nos centraremos en la descripción de tecnologías SBS. Encontramos múltiples tipos dentro de esta categoría SBS pero aquí tan solo revisaremos aquellos denominados de Adición Única de Nucleótidos (SNA, del inglés *"Single Nucleotide Addition"*) y los de Terminación del Ciclo Reversible (CRT, del inglés *"Cyclic reversible termination"*) [6] (Figura 4).

- **Adición Única de Nucleótidos (SNA)**: Esta metodología se caracteriza por detectar la señal emitida por cada nucleótido de forma independiente tras la adición única en cada ciclo sin necesidad de bloquear el proceso de elongación.

  La primera tecnología de secuenciación masiva que usaba este tipo de aproximación apareció en 2005 y fue la denominada *Roche 454* [29]. Está basada en la amplificación por *beads* (Box 4a) y distribución de estos a los largo de unos pocillos de reacción [55]. La obtención de la señal se produce mediante la pirosecuenciación (Figura 4a), es decir, por la emisión de luz acoplada a la reacción de elongación por parte de una luciferasa y detectada a través de un sensor acoplado. Supuso la primera revolución en el mercado de la secuenciación, disminuyendo el coste por base secuenciada e incrementando el rendimiento (Figuras 2-3).

- **Terminación del Ciclo Reversible (CRT)**: Unos años más tarde, hacia 2007, aparecieron las tecnologías basadas en la utilización de nucleótidos modificados (NRT, del inglés *"Nucleotide Reversible Terminators"*) (Box 1). El carbono 3' de la ribosa contiene un grupo bloqueante (R) que previene la

elongación [56, 57]. Además, cada base nitrogenada contiene un fluorocromo (F) que permite identificar cada tipo de base incorporado. Aunque estos análogos químicos de los nucleótidos son reconocidos por la polimerasa e incorporados, es necesario eliminar tanto el grupo R como el F para continuar la elongación de la cadena. Por tanto, mediante fluorescencia y previa rotura se puede detectar el nucleótido incorporado para después continuar con la elongación y el proceso de secuenciación (Figura 4b).

Como ejemplo de esta tecnología CRT encontramos la plataforma de secuenciación más empleada a nivel mundial de la empresa *Illumina*. Actualmente domina el mercado tanto por su precio, versatilidad y rango de plataformas con diferente rendimiento [4].

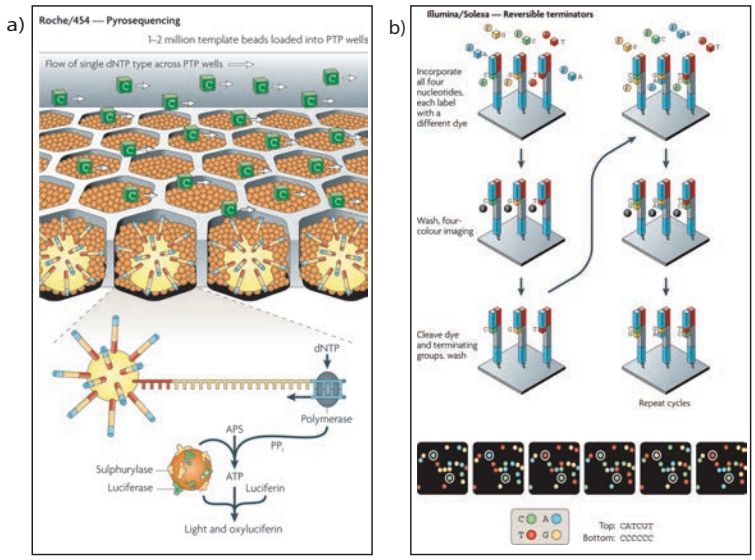
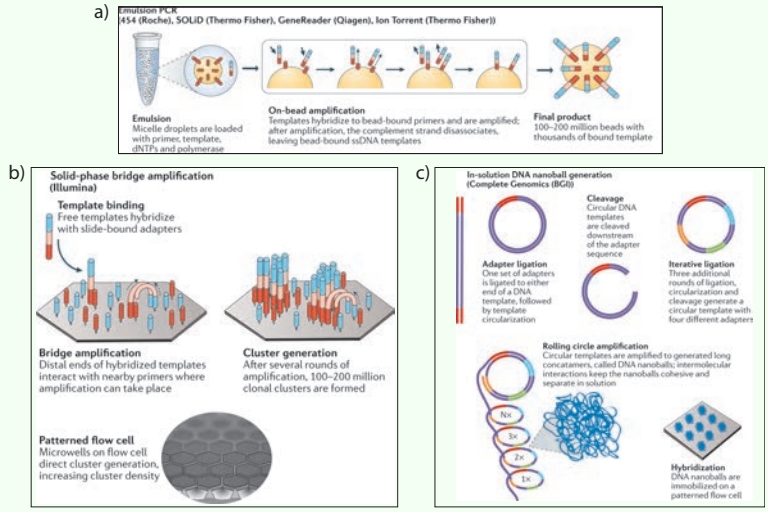


Figura 4: **Metodología usada por las principales técnicas de secuenciación masiva de nueva generación.** a) *Roche 454*; b) *Illumina*; Figura adaptada [4, 6].

Box 4: Técnicas de fijación y amplificación del ADN.



Encontramos múltiples técnicas para fijar a diferentes superficies y amplificar el ADN previo a la secuenciación. Aquellas que han tenido una mayor relevancia han sido:

(a) **Preparaciones basadas en *beads***: Los *beads* contienen adaptadores fijados para la amplificación y secuenciación y mediante una PCR (*"emulsion PCR"*) [58] se amplifican los fragmentos millones de veces en cada *bead* [5]. Una vez amplificados, se disponen distribuidos en una superficie [55, 59] para su posterior secuenciación. Un ejemplo de plataforma de secuenciación que utiliza esta metodología es la de *Roche 454*.

(b) **Amplificación en estado sólido**: Se evita el uso de una *"emulsion PCR"* al amplificar directamente los fragmentos en una superficie [60, 61]. Los *primers* están fijados de forma covalente a ésta y permiten la unión del fragmento de ADN. Un control preciso de la concentración de *primers* permite una distribución espaciada correcta para la secuenciación. También se pueden emplear superficies con celdas diseñadas que permitan mayor densidad y por tanto mayor rendimiento de secuenciación. Un ejemplo de plataforma de secuenciación que utiliza esta metodología son *Illumina* & *SOLiD*.

(c) **Enriquecimiento ADN en solución de nanopartículas**: Tan solo la compañía *Beijing Genomics Institute* (BGI) en su tecnología *Complete Genomics* utiliza esta metodología de fijación y amplificación del ADN. Básicamente, el ADN sufre un proceso iterativo de ligación, circularización y rotura para crear una solución de nanopartículas con ADN circular fijado para su posterior amplificación en una superficie [62].

Se puede apreciar cómo la aparición de este tipo de tecnologías redujeron considerablemente los costes por base secuenciada, aumentando el rendimiento de la secuenciación (Figuras 2-3). La aparición y popularización de este tipo de tecnologías entre la comunidad científica pronto se reflejó en los artículos científicos publicados. Se puede apreciar el incremento del número de artículos indexados en *PubMed* [10] sobre secuenciación y ensamblaje de genomas a partir del momento en el aparecen en el mercado este tipo de tecnologías de secuenciación masiva [63].

1.1.3. Nueva Secuenciación masiva: *"Third-generation sequencing"*.

Las metodologías citadas en el apartado 1.1.2 son capaces de generar *reads* cortos (de hasta unos 300 pb) (Figura 3), precisan de un proceso de amplificación con la consecuente inclusión de errores, generación de sesgos en la secuenciación y pérdida de información además del tiempo y coste añadido por los procesos de amplificación. Debido a la complejidad inherente del propio genoma como las repeticiones, alteraciones en el número de copias o las variantes estructurales, seguían existiendo cuestiones técnicas sin resolver que impedían obtener una secuencia genómica a nivel cromosómica.

Desde hace unos años varias empresas habían empezado a desarrollar metodologías de secuenciación de fragmentos largos a partir de una única molécula (SMS, del inglés *"Single Molecule Sequencing"*). A día de hoy encontramos dos tecnologías principales que de nuevo tras su aparición revolucionaron el mundo de la secuenciación masiva permitiendo obtener secuencias de muy buena continuidad y calidad. Son la tecnología desarrollada por *Pacific Biosciences* (*PacBio*) y la de *Oxford Nanopore Technologies* (ONT) o simplemente *Nanopore*.

La primera metodología desarrollada por *PacBio* [64] es la denominada secuenciación en tiempo real de una única molécula (SMRT, del inglés *"Single Molecule Real Time"*). Está basada en la actividad de síntesis de una ADN polimerasa modificada capaz de generar fragmentos de unos 10-15 kb (Figura 5a). Mientras que en la metodología SBS el ADN es anclado en una superficie y la ADN polimerasa está en disolución, en esta tecnología *PacBio*, la enzima se encuentra fijada al fondo del pocillo. La molécula de ADN debe entrar al pocillo y debido al reducido tamaño de éste tan sólo una única molécula podrá entrar

de forma simultánea. La lectura se realiza al incorporar cada nucleótido marcado evitando el uso de la amplificación.

Por otra parte, encontramos la segunda tecnología de secuenciación de fragmentos largos, *Nanopore*. A diferencia de otras metodologías, no se monitoriza la incorporación de nucleótidos ni se emplea una señal secundaria como la liberación de luz, protones o pH. Esta tecnología se caracteriza por la identificación de la composición de la molécula de ADN al atravesar un canal anclado en una membrana (Figura 5b). Este concepto fue hipotetizado por Deamer en 1989 y fueron necesarios muchos avances para convertir la teoría en realidad hacia el año 2010 [65, 66].

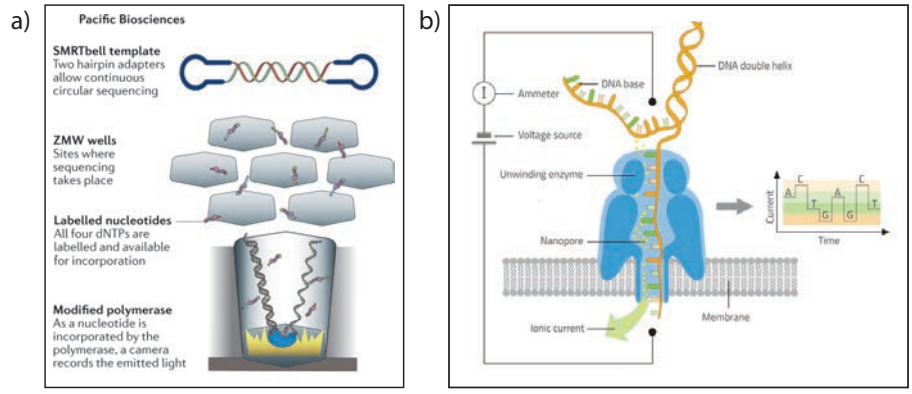


Figura 5: **Metodología usada por las principales técnicas de secuenciación masiva de tercera generación.** a) *Pacbio*. b) *Nanopore*. Figura adaptada [4, 6].

A pesar de que ambas tecnologías generan secuencias de una longitud similar, encontramos una gran ventaja de ONT respecto a *PacBio*, una longitud media de secuencia mayor, su rendimiento, el precio por base secuenciada y portabilidad (Figura 3; Tabla S2). ONT presenta múltiples plataformas de diferente tamaño y con distintos rendimientos, adaptándose a las necesidades de los usuarios. Encontramos desde plataformas de secuenciación que caben en un dispositivo de reducido tamaño y similar a una memoria *flash USB*, con el objetivo de ser empleadas en campañas de campo en zonas con pocas infraestructuras, hasta grandes máquinas de secuenciación con unos rendimientos muy altos.

El coste por base secuenciada de estas metodologías de *reads* largos, es mayor que en otras tecnologías como *Illumina*. Además, estas nuevas metodologías aún

se encuentran en fase de desarrollo y presentan una tasa de error de secuenciación mayor que las citadas en el apartado 1.1.2. Pero a su vez, presentan múltiples ventajas que hacen de esta tecnología una metodología muy prometedora [67]. Estas metodologías de secuenciación de fragmentos largos son capaces de generar lecturas de varias kb permitiendo así atravesar y resolver zonas altamente repetitivas o con características estructurales complejas. Por lo general mediante el uso de estas tecnologías se consiguen ensamblajes mucho más continuos y completos. Estos *reads* largos también permiten una determinación clara de isoformas al estudiar un transcriptoma ya que un único fragmento se expande por todo el transcrito permitiendo conocer las diferentes conexiones de exones.

Además de las tecnologías de SMS, encontramos otras tecnologías desarrolladas y englobadas dentro de la secuenciación de nueva generación. Pondremos como ejemplo, los métodos de captura de la conformación de los cromosomas, los mapeos ópticos y la secuenciación de células únicas.

La técnica denominada captura de la conformación cromosómica (del inglés *"genome-wide conformation capture"*) permite determinar la organización espacial de cromatina [68]. Este método cuantifica el número de interacciones entre *loci* genómicos próximos en el espacio. Las frecuencias de interacción pueden ser analizadas directamente o pueden ser utilizadas para reconstruir estructuras *3-D*. Para realizar esta técnica es necesario determinar previamente la interacción a nivel molecular. Para ello se añade formaldehído que conservará las interacciones y una endonucleasa de restricción con una diana de 4-6 pb que permita, de forma aleatoria, fragmentar el genoma en numerosos fragmentos. A continuación se produce la ligación de los fragmentos y su posterior secuenciación atendiendo al tipo de análisis. Encontramos variantes de esta metodología que permiten [68]: identificar la interacción entre dos fragmentos concretos (3C) mediante *primer*s de PCR; conocer la interacción de un único fragmento con el resto del genoma (4C) mediante *chips* de secuenciación; o conocer la interacción de cualquier fragmento que ocurra en todo el genoma (Hi-C, del inglés *"High-throughput sequencing Genome-wide chromosome conformation capture"*) mediante tecnologías de secuenciación masiva. Por otra parte, variantes de estas metodologías permiten enriquecer la muestra en solo unas regiones de interés o conocer la interacción con determinadas proteínas de la cromatina.

Las aplicaciones de este tipo de análisis de la conformación cromosómica son muchas y han permitido conocer por ejemplo cómo interaccionan los factores de

transcripción con los genes que regulan. También han permitido aumentar la continuidad de los ensamblajes al poder ordenar en *scaffold*s los diferentes contigs ensamblados [69]. Una de las empresas más conocidas que ha incorporado esta técnica de conformación cromosómica a la mejora de los ensamblajes genómicos, es la empresa *Dovetail genomics*. A partir de un ensamblaje inicial esta empresa es capaz de mejorar la continuidad de este ensamblaje mediante secuenciación tipo Hi-C, denominada Chicago y su posterior análisis bioinformático. Se consigue aumentar en más de 2-3 órdenes de magnitud la continuidad del genoma ensamblado generando secuencias a nivel cromosómico [70].

Otra técnica es la de mapeo óptico de una molécula de ADN que permite añadir un contexto a la secuencia que falta en la secuenciación de fragmentos de ADN cortos. Mediante un mapeo óptico se consigue marcar cada cierto número de kb una molécula de ADN de una forma determinada y conocida. La modificación específica y controlada del ADN, fijado en una superficie, se produce mediante fluorescencia o mediante modificaciones enzimáticas (*e.g.* enzimas de restricción, metiltransferasas, etc). Tras el marcaje se produce un procesamiento y análisis de la imagen generada. Este mapeo óptico complementa la información obtenida mediante secuenciación y ensamblaje al generar patrones de distribución de las regiones marcadas que ayudan en el ensamblaje y generación de *scaffolds* [71]. El primer genoma analizado mediante este tipo de aproximaciones fue el de *Sacharomyces cerevisae* (1993) [72]. En la actualidad se ha utilizado en la mejora de la continuidad de ensamblajes genómicos de especies con genomas grandes y repetitivos (*e.g.* maiz [73]) y en el estudio de de variantes estructurales largas en humanos [74]).

Por último encontramos las tecnologías de secuenciación de células individuales (scSeq, del inglés *"Single-cell sequencing"*). Este tipo de metodologías permiten obtener información a partir de una única célula partir de ADN, ARN o información de la metilación del ADN, o de la conformación cromosómica de la célula. El proceso comienza con la separación de cada célula, generalmente mediante microfluidos, en pequeños capilares donde se produce la extracción del material genético de interés y la preparación de la librería de secuenciación para cada célula. Mediante el uso de robots se consigue obtener un rendimiento muy alto de este proceso y generar una secuenciación de varias miles de células.

1.2. La era *"Ómica"* en biología

En el apartado 1.1 se ha puesto de manifiesto cómo la historia de la secuenciación está repleta de diferentes hitos que han revolucionado el mundo científico y el acceso y popularización de la secuenciación por parte de la comunidad científica. Por una parte se ha visto cómo en los primeros años de la historia de la secuenciación sucedieron numerosos avances técnicos. A continuación la aparición de diferentes tecnologías de secuenciación masiva incrementaron la cantidad de información obtenida. Fueron necesarios avances en paralelo de la tecnología, la informática y la aplicación al mundo de la biología, la bioinformática. Importantes proyectos como el proyecto HGP sirvieron de catalizador de todos estos avances. Aunque seguían existiendo limitaciones técnicas, las preguntas biológicas que se podían intentar contestar crecían exponencialmente. En los últimos años, avances en metodologías de *reads* largos permitieron obtener mucha más y continua información y solventar así problemas de otras metodologías. Otro tipo de metodologías también desarrolladas últimamente han permitido analizar y resolver cuestiones biológicas que hasta hace unos años resultaban imposibles, como la heterogeneidad de tipos celulares en el cáncer [75], la diversidad de especies en determinados ambientes [76] ó la evolución de líneas celulares [77] entre otras cuestiones.

Por tanto, la importancia de la secuenciación de ácidos nucleícos es una cuestión indiscutible hoy en día en el mundo de la biología tanto básica como aplicada. Algunos autores han indicado que las técnicas de secuenciación masiva podrían tener tanta repercusión en el último siglo como el desarrollo del microscopio [5] o la biología molecular, o el impacto de la teoría neodarwinista [78].

Como hemos visto anteriormente, el auge, la popularización y la aplicación de la secuenciación masiva hizo necesario el desarrollo en paralelo de la bioinformática (Box 2) que se puede apreciar en la figura 2. Se puede observar cómo ha aumentado en los últimos años el número de artículos indexados en *PubMed* que contienen términos relacionados (NGS, secuenciación de genoma, bioinformática). También se aprecia por el incremento de secuencias incorporadas en la base de datos *GenBank* o por la cantidad de especies eucariotas cuyo genoma ha sido secuenciado e incorporado en la base de datos *GenBank* (8942) (Tabla 1) [8, 79].

(a) Animales

Anfibios	9
Insectos	626
Mamíferos	722
Nematodos	177
Pajaros	187
Peces	321
Platelmintos	50
Reptiles	49
Otros animales	226
Total	2367

(b) Plantas

Algas verdes	89
Plantas terrestres	911
Otras plantas	3
Total	1003

(c) Hongos

Ascomicetos	3807
Basidiomicetos	801
Otros fungi	213
Total	4821

(d) Protistas

Apicomplejos	266
Cinetoplastos	118
Otros protistas	367
Total	751

Tabla 1: **Genomas eucariotas secuenciados hasta la fecha e indexados en la base de datos *GenBank***. Clasificados para cada subgrupo. Datos obtenidos a fecha de Septiembre 2019 [8, 79].

Los diferentes hitos sucedidos en la historia de la secuenciación y de la bioinformática han permitido un gran incremento del ámbito de aplicaciones de estas metodologías de secuenciación masiva (Box 5). En cuanto a la diferente aplicación de estas tecnologías en el contexto de esta tesis, a continuación tan solo revisaremos brevemente diversas aplicaciones.

1.2.1. Secuenciación de genomas: ensamblajes *de novo*.

El motor que incentivó el desarrollo y evolución de las metodologías de secuenciación masiva, y en paralelo de la bioinformática, fue la gran utilidad de la información proporcionada por esta secuencia genómica obtenida tanto para la biología básica como aplicada. Como hemos visto, numerosos hitos, tanto a nivel tecnológico, informático o técnico se tuvieron que conseguir para lograrlo (Figura 1).

Con la aparición en 2005 de las metodologías NGS, el número de genomas secuenciados e incorporados en las bases de datos aumentó considerablemente (Figura 2). Debido a las limitaciones de las técnicas de secuenciación, a los sesgos de los procesos de amplificación y secuenciación y el gran tamaño del genoma de algunos organismos, existían regiones que no se podían ensamblar por su complejidad. Fue necesario el desarrollo paralelo de técnicas que solventasen estos problemas como la secuenciación pareada (PE, del inglés *"paired-end"*) [80] con tamaños de insertos de menor o mayor longitud (MP, del inglés *"mate pair"*) y secuenciación de *reads* largos (SMS) capaces de expandir regiones repetitivas (*PacBio*, *Nanopore*).

Box 5: Tipos de tecnologías *"Ómicas"*

Las aplicaciones de NGS han permitido el estudio del conjunto en muchos ámbitos de la biología molecular y ha ido evolucionando en concordancia con las limitaciones en cada momento.
Encontramos multitud de ejemplos de estas tecnologías *"ómicas"*:

- Genómica: secuenciación del genoma completo (WGS, del inglés *"Whole genome sequencing"*) o de exones (en inglés *"exome sequencing"*).
- Transcriptómica: secuenciación de transcritos o RNAseq [81].
- Metagenómica: secuenciación de muestras ambientales para la identificación de la colección de microorganismos presentes.
- Proteómica: secuenciación a gran escala de proteínas para el conocimiento de su estructura y función.
- Metabolómica: secuenciación del conjunto de metabolitos presentes en una célula, tejido, muestra ambiental, etc.
- Epigenómica: estudio de las modificaciones de la cromatina. Encontramos un tipo de secuenciación para el estado compacto de la cromatina, ChIP-seq (del inglés *"Chromatin immunoprecipitation followed by sequencing"*) [82], para el estado abierto y accesible, ATAC-seq (del inglés *"Assay for transposable accessible chromatin using sequencing"*) [83] y otro para los patrones de metilación, methyl-seq (del inglés *"DNA methylation sequencing"*).

En la actualidad, las estrategias empleadas más comúnmente para la secuenciación *de novo* de genomas grandes, complejos y con una gran número de repeticiones, consisten en combinar una cantidad alta de *reads* de *Illumina* PE con *reads* MP de diferente tamaño de inserto (últimamente en desuso) y con una cantidad moderada de *reads* largos que permitan a los diferentes *softwares* bioinformáticos (*e.g.* MaSuRCa [84], ALL-PATHS [85]) realizar ensamblajes híbridos aprovechando las características de los diferentes tipos de secuenciación. Adicionalmente, el empleo de mapeos ópticos ó Hi-C permitiría mejorar considerablemente la continuidad del genoma.

1.2.2. Anotación estructural y funcional de genomas.

Una vez obtenida la secuencia ensamblada es necesario proceder a la identificación y caracterización de los genes que conforman y caracterizan a ese genoma. El proceso de anotar incluye la identificación de regiones codificadoras (CDS, del inglés *"Coding sequence"*) y no codificadoras (UTRs, del inglés *"Untranslated regions"*) [86]. Una buena metodología de anotación debe incluir una predicción *de novo* de CDS y UTRs, así como una anotación basada en evidencias biológicas de transcritos y de proteínas homólogas.

Actualmente existe una gran cantidad de genomas secuenciados, anotados e indexados en la base de datos *GenBank* [8, 79]. En la tabla 1 se indican sólo el número de genomas de especies distintas, ya que además, para muchas especies existen proyectos donde se ha resecuenciado el genoma de varios individuos (*e.g.* estudios de genética de poblaciones). Los primeros genomas eucariotas secuenciados fueron principalmente organismos modelo como *Saccharomyces cerevisiae* [87] (1996), *Caenorhabditis elegans* [88] (1998), *Arabidopsis thaliana* [89] (2000), *Drosophila melanogaster* [90] (2000), *Homo sapiens* [46, 53] (2001) y *Mus musculus* [91] (2002) (Figura 1). Viendo la cantidad y diversidad de genomas secuenciados a día de hoy (8942) es indudable que la mayoría representan ya especies no modelo [78].

Debido al interés económico, social o sanitario, la lista de organismos con genoma secuenciado está sesgada hacia ciertos grupos taxonómicos. Por ejemplo $>0.1\,\%$ de las especies de vertebrados identificadas han sido secuenciadas, donde los mamíferos representan el grupo mejor caracterizado hasta la fecha, con $>1\,\%$ de representantes secuenciados. Por otra parte, la proporción de especies secuen-

ciadas (sobre las identificadas) en otros grupos es varios órdenes de magnitud inferior; por ejemplo para plantas y hongos en torno a un ~0.01 % mientras que para insectos mucho inferior (~0.001 %). Además, la lista de eucariotas secuenciados está sesgada hacia especies domesticadas de interés agronómico, horticultural o hacia especies de importante relevancia ecológica o evolutiva [78].

Toda esta variedad de genomas disponibles está cambiando la forma de abordar cuestiones tanto de la biología básica [78, 92–94] como de la aplicada, incluyendo la medicina y clínica [5, 95, 96]. Muchos de los proyectos de secuenciación genómica de organismos no modelo han llevado a la identificación de genes específicos de linaje y a expansiones/contracciones de familias multigénicas con mayor o menor repercusión fenotípica y evolutiva [78] (Tabla 2). Encontramos también algunos ejemplos de proyectos genómicos y de sus principales aportaciones en los que ha participado el grupo de investigación donde se ha realizado esta tesis doctoral (Tabla 3).

1.2.3. Análisis genómico evolutivos en organismos no modelo: desarrollo de marcadores moleculares.

Como veíamos antes (apartado 1.2.2), encontramos una gran cantidad de información genómica disponible para un amplio rango taxonómico (Tabla 1), principalmente para especies modelo pero también de organismos no modelo [106]. La capacidad de acceder a la información genómica completa de múltiples especies ha permitido realizar estudios de procesos evolutivos y de evolución molecular usando datos genómicos en vez de unos cuantos marcadores. Así se han podido abordar cuestiones evolutivas, de conservación o ecológicas usando datos a nivel genómico [94, 107].

A pesar de los múltiples avances en la secuenciación genómica, aún no se ha llegado al punto en que de forma rutinaria y eficiente seamos capaces de generar un ensamblaje genómico continuo de calidad en organismos eucariotas [108]. Para muchos grupos de investigación y a pesar de la reducción de costes de secuenciación, o no necesitan o no son capaces de realizar este tipo de secuenciación y procesamiento de los datos. Prefieren seguir usando un conjunto limitado de marcadores para los cuales disponen de información previa o de especies cercanas o de cierto interés.

Tabla 2: **Ejemplos de principales descubrimientos en algunos proyectos de secuenciación de genomas de organismos.**

Nombre común	Nombre científico	Descubrimiento	Referencias[a]
Carbonero terrestre	*Pseudopodoces humilis*	Expansión de familias multigénicas implicadas en el metabolismo y relacionadas con el estilo de vida en altura	[21]
Mariposa Postman	*Heliconius melpomene*	Expansión de genes del sistema quimiosensorial. Complejidad visual facilitada por la expresión de una opsina ultravioleta duplicada	[3]
Ostra del Pacífico	*Crassostrea gigas*	Expansión de genes inhibidores de la apoptosis y *"heat shock protein"* involucrados en la protección frente al calor y estrés	[13]
Trigo	*Triticum aestivum*	Expansión de familas génicas, como resultado de la domesticación de este vegetal, relacionadas con la defensa, contenido nutricional, metabolismo y crecimiento	[2]
Tomate	*Solanum lycopersicum*	Ejemplos de neofuncionalización debido a eventos de triplicaciones genómicas. Expansiones de genes involucrados en la modificación de la pared y del desarrollo del fruto	[4]
Algodón	*Gossypium raimondii*	Alta complejidad genética debido el aumento de 5-6 veces de ploidia y a una posterior reducción	[104:105]
Naranja	*Citrus sinensis*	Especie derivada de un retrocruzamiento mediante el pomelo y la mandarina	[106]
Melocotón	*Prunus persica*	Expansión de genes relacionados con el metabolismo del sorbitol que han proporcionado el sabor dulce	[107]

[a]Para más información sobre cada proyecto acudid a la referencia indicada en Ellegren, 2014 [78].

Tabla 3: **Proyectos de secuenciación de genomas donde ha participado el grupo de investigación.**

Nombre común	Nombre científico	Descubrimiento	Referencias
Piojo	*Pediculus humanus humanus*	Ejemplo de miniaturización del genoma de un endosimbionte y de las relaciones con el hospedador y posibles patógenos.	[97]
Pulgón	*Acyrthosiphon pisum*	Importantes ganancias y perdidas de genes; interacción con simbiontes	[98]
Planta de cafe	*Coffea canephora*	Familias multigénicas especificas involucradas en la produccion de cafeína, alcanoides y flavonoides y genes de defensa. Posible mecanismo de expansión mediante duplicaciones en tandem de estos genes.	[99]
Ciempiés	*Strigamia maritima*	Identificación de la aparición de algunos mecanismos y rasgos a los largo de la evolución de los artrópodos.	[100]
Planta carnívora	*Utricularia gibba*	Familias génicas relacionadas con la adaptación funcional, el estilo de vida y arquitectura de la planta	[101]
Garrapata	*Ixodes scapularis*	Genes importantes con el estilo de vida parasítico.	[102]
Polilla	*Manduca sexta*	Anotación detallada de genes y familias multigénicas, tanto a nivel estructural como funcional, relacionadas con la posible aplicación en control de plagas y como organismo modelo	[103]
Cephalotus	*Cephalotus follicularis*	Selección positiva de genes relacionados con el modo de vida, digestión de presas, genes de respuesta a estrés, etc.	[104]
Aguacate	*Persea americana*	Grandes duplicaciones genómicas; historia evolutiva reciente ligada a la interacción con patogenos	[105]
Filoxera	*Dactylosphaera vitifoliae*	Caracterización de genes involucrados en su modo de vida y relación con el hospedador, con potencial para el control de plagas.	*In preparation*

La proliferación de métodos de partición genómica y técnicas para el enriquecimiento de librerías de secuenciación [108, 109] de restringidas regiones genómicas de interés, han supuesto un importante avance en sistemática o en filogenia molecular [94]. Mediantes estos métodos, se puede secuenciar tan solo unas pocas regiones genómicas, de las cuales existe normalmente información previa, reduciendo la cantidad y el coste de secuenciación, así como el posterior procesamiento de los datos. De esta manera, se puede aumentar el número de organismos que se pueden estudiar (secuencias), aumentando la efectividad y el potencial de resolución filogenético del estudio.

A continuación indicamos algunas estrategias de partición genómicas aplicables a estudios filogenéticos o de genética de poblaciones:

- **Amplificación de fragmentos dirigida** (TAS, del inglés *"Targeted Amplicon Sequencing"*) [110]. Consiste en la amplificación de diferentes fragmentos (marcadores) a lo largo del genoma con *primers* específicos para su posterior etiquetado mediante marcas de nucleótidos para cada especie y su posterior secuenciación masiva.
- **Librerías reducidas representativas** (RRL, del inglés *"Reduced representation library"*) [111]. Consisten en la digestión del genoma mediante enzimas de restricción para obtener un conjunto de fragmentos, en principio aleatorio, del genoma. Encontramos asociada a esta reducción del genoma una técnica de secuenciación (RADseq, del inglés *"Restriction-site-associated DNA sequencing"*) [112, 113] mediante la cual tan solo aquellas bases adyacentes al sitio de restricción son secuenciadas.
- **Análisis del transcriptoma** de un organismo mediante RNAseq: es una muy buena estrategia aplicable a estudios filogenéticos [81, 94] ya no solo por la generación de secuencias codificantes sino también por las secuencias generadas por *splicing* alternativo.
- **Captura de secuencias** denominada (del inglés *"Hybrid Enrichment"*). Ésta implica la síntesis de oligonucleótidos complementarios a las regiones de interés del genoma. Mediante un proceso de hibridación en disolución [114] se capturan las regiones de interés y se procede a la secuenciación. Dentro de este grupo encontramos los denominados UCEs (del inglés *"Ultraconserved Elements"*) [115] que aportan diferentes niveles de variabilidad y permiten ser aplicados en un amplio rango de grupos taxonómicos.

Estas estrategias tienen una aplicabilidad que difiere según el rango taxonómico que se quiera analizar, el tamaño del genoma y la proporción de este que se quiere particionar. Cada una presenta múltiples ventajas y desventajas (Box 6) lo que genera un abanico amplio de posibilidades para el investigador que dependerá del problema biológico que quiera abordar.

Desde hace años se conocen los problemas que pueden generar estudios filogenéticos de pocos marcadores [116] o aquellos basados en solo marcadores mitocondriales [117]. La disponibilidad de múltiples marcadores distribuidos a lo largo del genoma de origen independiente, no ligados y de procedencia tanto mitocondrial como nuclear resulta crucial para una reconstrucción filogenética fiable [118, 119]. Debido a la falta de información genómica de la especie de interés o de alguna cercana [106], en algunos estudios resulta costoso el disponer de múltiples marcadores.

Encontramos diversas aproximaciones para identificar marcadores moleculares en organismos no modelo, con sus ventajas y desventajas (Tabla 4). Elegir la estrategia adecuada es un equilibrio entre las características deseadas, la disponibilidad y el tipo de material genético así como del presupuesto disponible. Los marcadores tipo NPCL (del inglés *"Nuclear Protein Coding Loci"*) [120, 121] y los tipo EPIC (del inglés *"Exon Primed Intron Crossing"*) [122] dependen de la disponibilidad de un genoma anotado de una especie cercana. Utilizando la anotación disponible el desarrollo de marcadores es directo, rápido y dirigido hacia zonas codificantes o no codificantes y evitando zonas con repeticiones o posibles parálogos [107].

Por otra parte, encontramos los UCEs [115], basados en métodos de enriquecimiento de secuencia y aunque no dependen de un genoma secuenciado, es necesario haberlos diseñado previamente en un rango taxonómico que incluya la especie de interés. Otros marcadores son los asociados a una reducción del genoma por RRL como los de RADseq [112, 113]. Estos métodos han sido aplicados en un gran número de proyectos de genotipado y genómica de poblaciones. Por el diseño de la técnica, tan solo unas pocas bases adyacentes al sitio de restricción son secuenciadas y por tanto, en general, no se pueden realizar análisis de marcadores ligados. Además, existe una limitación para su aplicabilidad debido a las sustituciones nucleotídicas en las dianas de restricción que conlleva que no siempre se pueda utilizar en un grupo amplio de especies. Para un mayor detalle de las características de RADseq se pueden revisar otros artículos [94, 123, 124].

Box 6: Características de las estrategias de reducción genómica.

Se indican varias caracaterísticas entre las que destacamos algunas de ellas. Los círculos azules indican que para ese criterio el método funciona correctamente, mientras si se encuentra a medias o vacío, tiene un rendimiento moderado o bajo, respectivamente. Se incluye la secuenciación genómica (WGS) para ilustrar la capacidad límite de esa técnica. Fuente: [94]

- **Aplicación en organismos:** *"Broad taxonomic application"*, sí un esfuerzo inicial es suficiente para su aplicación a un rango profundo filogenético; *"Efficient with non model species"*, sí se puede directamente aplicar a organismos sin ningún conocimiento previo; *"Usable across Tree of Life"*: si se puede emplear en cualquier grupo taxonómico. *"Broad phylogenetic informativeness"*: si se puede emplear a cualquier escala filogenética.
- **Tipo de marcadores:** *"Long loci"*, si permite obtener fragmentos largos (>1 kb); *"Utilizes DNA"*, indica el tipo de ácido nucleico que se puede emplear: ADN (círculo lleno) o ARN (círculo vacío). *"Accomodates degraded samples"*, si requiere material genético de muy alta calidad.
- **Proceso:** *"Short development time"*, el tiempo relativo de desarrollo del proyecto inicial. *"Short data collection time"*, el tiempo relativo para producir información filogenética a partir de este método. *"Short data assembly time"*, el tiempo relativo de procesamiento bioinformático de los resultados obtenidos. *"Cost effective"*, la ratio de efectividad del método.
- **Otras:** *"Minimal missing data"*, si obtiene datos para todas las muestras o se permite generar cierta pérdida de información; *"Robust to substitutions"*, la sensibilidad a los sustituciones de nucleótidos en las regiones amplificadas; *"Efficient with large genome"*, si es eficiente con genomas grandes.

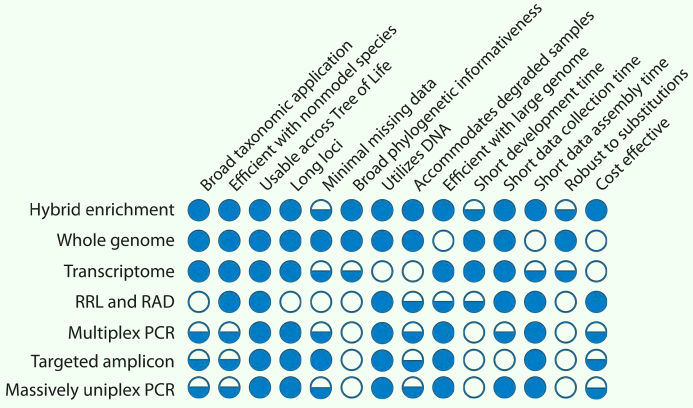


Tabla 4: **Principales ventajas e inconvenientes de los diferentes tipos de marcadores moleculares.**

Marcador	Ventajas	Desventajas
NPCL[a] y EPIC[b]	Desarrollo directo	Precisa de un genoma bien anotado de un organismo cercano
UCEs[c]	Apropiado para un amplio rango filogenético	Necesista secuencias identificados previamente y métodos de captura de secuencia
RADseq[d]	No información genómica previa. Gran cantidad de marcadores a bajo coste	Sesgado por polimorfismos en las dianas de restricción. No apropiado para estudio de ligamiento
ANMs[e]	No información genómica previa. Distribución aleatoria	Características (estructura, longitud, niveles de divergencaia) variables. Regiones anónimas.

[a]NPCL: *Nuclear Protein Coding Loci*; [b]EPIC: *Exon Primed Intron Crossing*; [c]UCE: *Ultranconserved Elements*; [d]RADseq: *Restriction site associated sequencing*; [e]ANM: *Anonymous molecular markers*;.

Por último encontramos los denominados marcadores moleculares nucleares anónimos (ANMs, del inglés *"Anonymous nuclear markers"*) [125]. Estos marcadores tienen una distribución relativamente uniforme a lo largo del genoma, aleatoria y en principio con una alta probabilidad de estar en regiones no codificantes [107]. No estarán ligados ni sesgados en una área del genoma y por tanto estarán sometidos a una evolución neutra que permite la acumulación de variación informativa para muchos estudios evolutivos [126]. Los marcadores desarrollados por esta técnica son adecuados para estudios a nivel filogenético, filogeográfico o a nivel de genética de poblaciones [94, 107, 119, 126].

Una manera muy eficiente de particionar y de muestrear el genoma, generando un conjunto de ANMs es mediante las técnicas RRL [111] y la posterior selección por tamaño. Entre los principales beneficios encontramos la reducción de los costes de secuenciación y del procesamiento de la información generada [127]. Al igual que ocurre en otras técnicas basadas en RRL, *e.g.* RADseq, una limitación

puede ser la pérdida de las dianas de restricción por divergencia interespecífica. Puesto que normalmente no existe información previa de la región a secuenciar, se pueden incluir parálogos y repeticiones [107]. Las regiones obtenidas presentaran una longitud, variabilidad y estructura variables en las diferentes especies de interés.

Una limitación de esta técnica radica en el procesamiento de los datos y el análisis bioinformático. Se precisa disponer de una metodología que permita diferenciar y clasificar los marcadores moleculares obtenidos, en si son adecuados para un estudio de secuenciación o no. Se deberían tener en cuenta una serie de características (longitud, variabilidad, estructura, rango de especies cubierto, etc) para poder seleccionar aquellas regiones más informativas en cada caso. Esta metodología permitiría el aprovechamiento de los avances y costes reducidos de la secuenciación masiva en áreas como la filogeografía o filogenia. Además, la versatilidad, facilidad de uso y accesibilidad de esta metodología bioinformática a la comunidad científica interesada aumentaría las posibilidades de uso de este tipo de aproximaciones permitiendo obtener ANMs en organismos no modelo.

1.3. Organismos de estudio: arañas del género *Dysdera*

Los organismos de estudio de esta tesis doctoral han sido principalmente organismos no modelo pertenecientes al grupo de los artrópodos, en concreto, se ha profundizado en el estudio de organismos del orden Araneae, las arañas.

1.3.1. Relaciones filogenéticas del género de estudio.

El grupo de los artrópodos constituye el filo más numeroso y diverso del reino animal. Incluye a animales invertebrados dotados de un esqueleto externo y apéndices articulados, entre los que encontramos a los insecto, crustáceos, miriápodos y quelicerados (Figura 6).

Las arañas se encuentran dentro del grupo de los quelicerados y son depredadores dominantes en múltiples ecosistemas terrestres. Es un grupo muy amplio y diversificado con una amplitud grande de nichos ecológicos. Existen más de 45.000 especies documentadas hasta la fecha [128]. Presentan interesantes aspectos evolutivos como la presencia de veneno y de la seda, la evolución del sistema quimiosensorial o junto con el resto de artrópodos los procesos de terrestralizacion (Box 7).

El principal grupo que aquí nos concierne, el género *Dysdera* (Figuras 6-7), se encuentra dentro de la familia Dysderidae (Synspermiata, Araneomorphae, Araneae). Este género contiene un total de 282 especies [128], más de la mitad de la diversidad de la familia. Estas arañas son cazadores nocturnos que no producen telas para cazar pero si capullos de seda donde se resguardan durante el dia debajo de piedras, cortezas de árbol, hojarasca etc [129, 130]. La distribución de la mayoría del género es principalmente en la cuenca mediterránea y las Islas Macaronésicas, Europa central y Oriente medio aunque encontramos casos como el de *Dysdera crocata* que tiene una distribución casi cosmopolita [131].

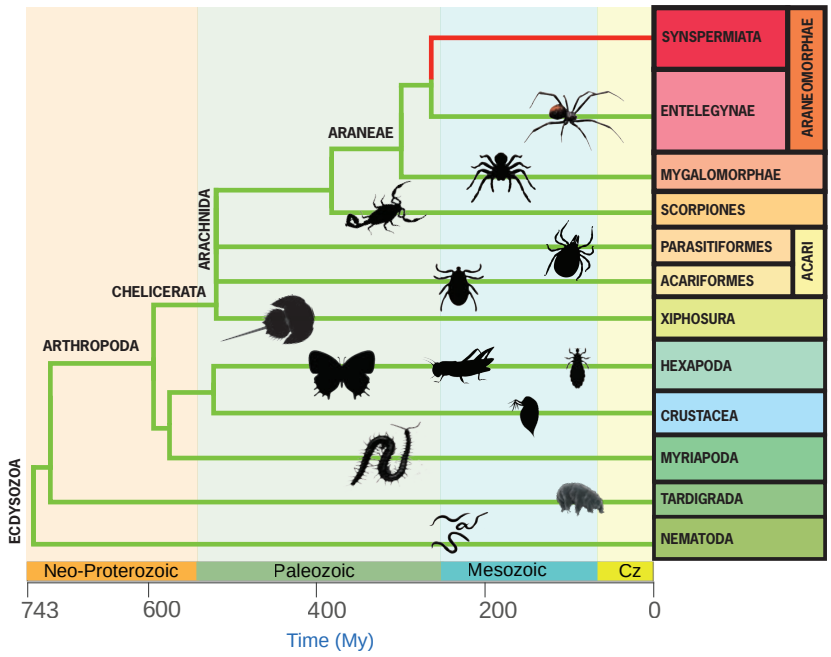


Figura 6: **Relaciones filogenéticas entres los grandes grupos de artrópodos**. En rojo encontramos el grupo de Sysnpermiata, donde se encuentra la familia Dysderidae y el género *Dysdera*. Arbol generado a partir de datos de Timetree [132] y tiempos de divergencia calculados según [133] y con los clados más conflictivos como politomias [134–136]. My, del inglés *"Million years"* y Cz, cretácico.

1.3.2. Radiación adaptativa y diversificación del género en las Islas Canarias.

Otro de los aspectos interesantes del género *Dysdera* es la increíble diversificación y radiación adaptativa en el archipiélago canario (Figura 8). Las Islas Canarias tienen un origen volcánico [137] debido a una pluma volcánica y numerosas fracturas de la corteza terrestre cuyo resultado actual es una graduación este - oeste en la edad de las islas siendo la más antigua, Fuerteventura, la más cercana al continente africano y con una edad cercana a 23 My (del inglés *"Millions years"*) (Figura 8a) y la más joven El Hierro con 1.1-1.2 My.

Box 7: Características biológicas relevantes del orden Araneae.

- **Seda**: Una de las características de especial interés económico es la producción de seda debido a las propiedades mecánicas y biomiméticas de sus fibras. A pesar de que la aparición de seda ha ocurrido varias veces, a lo largo de la evolución de los artrópodos, en las arañas ha alcanzado una gran sofisticación con hasta 12 tipos de fibras diferentes [138–140].
- **Veneno**: La producción de veneno, para capturar de presas y defensa, está presente en el grupo de los arácnidos casi la totalidad de escorpiones y arañas. Estos venenos presentan una gran complejidad y diversidad de compuestos moleculares, principalmente proteínas tóxicas. La presencia de este veneno también suscita una atención importante a nivel científico y social por sus posibles aplicaciones tanto en la biomedicina, biotecnología, control de plagas... [138, 141]
- **Duplicación genómica**: Los eventos de duplicaciones genómicas han sido los principales agentes de la diversificación y de la evolución de la complejidad de algunos grupos como los vertebrados [142]. En otros grupos como los cangrejos cacerola, representantes del grupo *Xiphosura* (Figura 6), no parece existir una correlación entre la duplicación genómica y la diversidad morfológica y genética [143]. Pero la duplicación genómica ocurrida en el grupo de las arañas datada tras la diversificación con los escorpiones supuso importantes eventos de neo-funcionalización y sub-funcionalización de algunos genes con importantes repercusiones en la diversidad y complejidad del grupo y la posible evolución de características específicas de estos organismos [140].
- **Colonización del medio terrestre**: La terrestralización se produjo de forma independiente para cada uno de los linajes del grupo de los artrópodos. Este proceso es clave en la diversificación y adaptación de las diferentes grupos. Se estima que este proceso se produjo en torno a 460 My [144] para los arácnidos tras comprobaciones con el registro fósil y con relojes moleculares.
- **Sistema quimiosensorial**: El sistema quimiosensorial es esencial para las principales funciones de los organismos pero mientras que en insectos está bien caracterizado desde hace años, en el resto de artrópodos, y en concreto en arácnidos, se desconocía. En los últimos años, los trabajos de Vizueta *et al.* determinaron que las familias multigénicas de los receptores gustativos (GR, del inglés *"gustatory receptors"*) e ionotrópicos (IR, del inglés *"ionotropic receptors"*) eran los candidatos principales quimiorreceptores en arañas. Identificaron una nueva familia denominada OBP de quelicerados (OBP-like del inglés *"odorant binding protein like"*) y otra familia denominada CCP (del inglés *"candidate carrier protein"*) [145] que junto con las proteínas NPC2 (del inglés *"Niemann–Pick protein type C2"*) estarían involucradas en el transporte de los ligandos a los diferentes receptores, con una función similar a las proteínas de insectos de unión a odorantes (OBP) o a las proteínas quimiosensoriales (CSP, del inglés *"chemosensory proteins"*).

Figura 7: **Fotografías de diferentes especies de arañas del género *Dysdera*.** a) *Dysdera verneaui*. b) *Dysdera tilosensis*. c) *Dysdera silvatica*, capturando un isópodo. d) *Dysdera gomerensis*. e) *Dysdera bandamae*. Fotografías: Pedro Oromí.

Las islas oceánicas han sido consideradas como laboratorios de la evolución y se consideran buenos sistemas modelo para la investigación de biogeografía, migraciones, diversificaciones y extinciones [146, 147]. En aquellos archipiélagos originados por plumas volcánicas se cumple la denominada *Ley de progresión* [146] que predice una concordancia entre la filogeografía de las especies y la edad geológica de las islas. Las islas oceánicas aportan casi un 20 % de la diversidad de especies terrestres con tan solo 3.5 % de toda la superficie emergida [147].

Encontramos numerosos ejemplos de radiaciones y endemismos de organismos en islas oceánicas. Por ejemplo en las Islas Canarias, existen ejemplos en plantas como el pino canario [148] y plantas herbáceas del género *Sonchus sp* [149]. Otros ejemplos documentados son en vertebrados como lagartos del género *Tarentola sp*, *Gallotia sp* y *Chalcides sp* [150–152]. En todo caso, el mayor número de ejemplos de radiaciones adaptativas en las Islas Canarias es de invertebrados: gasterópodos (*Napaeus sp*), isópodos (*Porcellio sp*) [152], quelicerados [153, 154], colémbolos, coleópteros, hemiptera, etc [152].

La diversificación del género *Dysdera* en las Islas Canarias ocurrió a partir de dos o tres eventos de colonización [155] desde el continente africano y su posterior diversificación incluso *in situ* [156] que dieron lugar a un total de 55 especies endémicas, muchas de ellas endémicas dentro de una isla [157] (Figura 8b, Tabla S4). Debido a las características de las diferentes islas como la presencia de volcanes de gran altitud y la localización geográfica [152], se puede encontrar una amplia variedad de nichos terrestres. Se ha detectado la presencia de *Dysdera* desde playas hasta grandes altitudes e incluso en ambientes subterráneos [156–159].

Encontramos múltiples casos de especies del género *Dysdera* que concurren en las mismas localidades pero que difieren en la forma y tamaño de quelíceros (Box 8) sugiriendo una especialización y adaptación a diferentes tipos de dieta [160, 161]. Se han realizado experimentos tanto de comportamiento como fisiológicos donde se ha observado una significativa asociación entre caracteres morfológicos y preferencia de presas. Un aspecto destacable de este género es que se considera un ejemplo de especialización trófica (estenofagia) en arañas, en este caso por isópodos. Estos isópodos presentan adaptaciones en el exoesqueleto, en el comportamiento y en técnicas de defensa, como la acumulación de metales pesados y toxinas [162–164], que los convierten en presas difíciles de atrapar y de digerir. Se conoce que las especies de isópodos que estas arañas capturan son endémicas de las islas y pertenecen al género *Porcellio sp*, u otras especies cosmopolitas e introducidas en las islas como *Armadillium vulgare* y *Eluma caelata*. Dentro del género *Dysdera* encontramos múltiples ejemplos de especies con capacidad y preferencia por la captura de isópodos y que presentan adaptaciones tanto metabólicas, morfológicas (quelíceros) (Box 8) como de comportamiento (estrategias de captura) [160, 161, 164–167]. Son las denominadas especies especialistas de dieta. Pero no todas las *Dysderas* se alimentan preferentemente de

isópodos, encontramos especies generalistas capaces de tomar cualquier otra presa con la misma preferencia, incluyendo isópodos, aunque también encontramos otras especies que tienden a evitarlos [161, 165, 167].

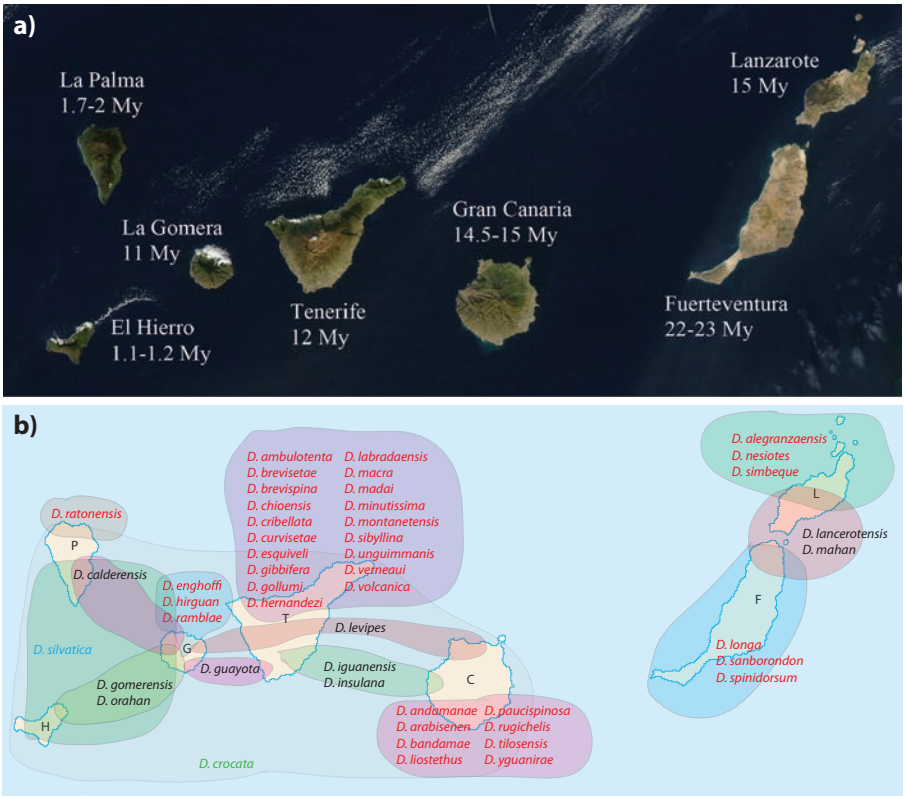
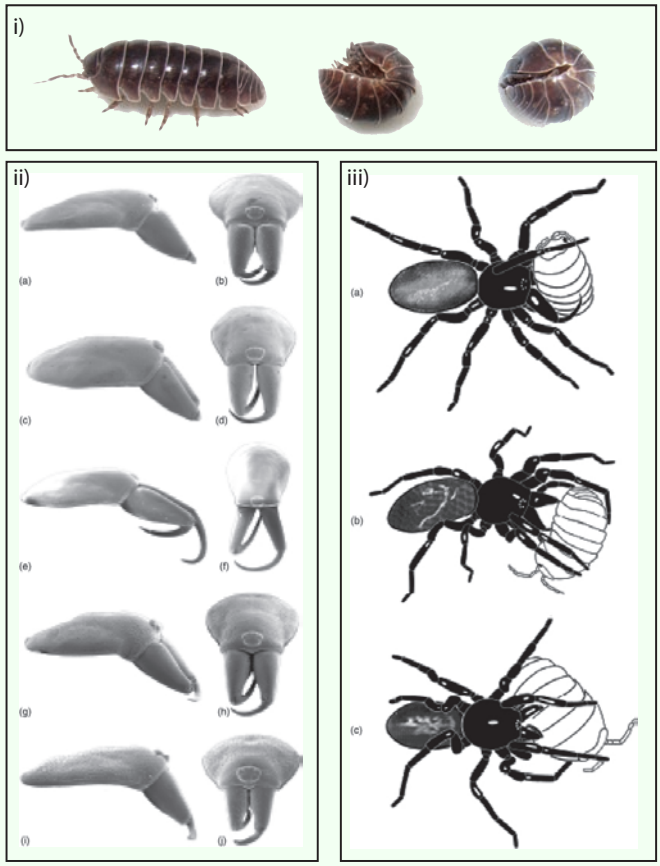


Figura 8: **Edad geológica de las Islas Canarias y distribución de las especies de *Dysdera* en el archipiélago**. a) Edad geológica expresada en My b) Distribución de especies del género *Dysdera* en las Islas Canarias representada mediante areas de diferentes colores (P: La Palma, H: El Hierro, G: La Gomera, T: Tenerife, C: Gran Canaria, F: Fuerteventura, L: Lanzarote). En rojo las especies endémicas para cada isla; en negro las especies endémicas del archipiélago presentes en múltiples islas; en azul *Dysdera silvatica*; en verde *Dysdera crocata*; Fuente [157]. Detalle en tabla S4.

Box 8: *Dysdera* sp. e isópodos.



i) Isópodo común (*Armadillium vulgare*). Estrategia de defensa para proteger sus patas y antenas. ii) Vista frontal y lateral de quelíceros de diferentes especies de arañas del género *Dysdera* [161]. Encontramos quelíceros (a y b) no modificados, otros elongados ligeramente (c y d) o muy elongados (e y f) y otros con forma concava (g y h) o aplanados (i y j). iii) Tácticas de sujeción y ataque de *Dysderas* a isópodos [161]. Encontramos tácticas de pinza (a), empleados por aquellas con quelíceros elongados; tácticas de trinchar (b), por aquellas con quelíceros cóncavos y tácticas de llave (c) por aquellas con quelíceros aplanados.

Las líneas de investigación del grupo donde se ha desarrollado esta tesis doctoral estan relacionadas con la comprensión de la evolución molecular y el estudio de procesos y mecanisnos evolutivos utilizando la genómica y transcriptómica. El organismo de estudio o género en este caso, *Dysdera sp*, contiene las características necesarias para convertirse en un buen candidato, como especie modelo, para el estudio de la determinación de la base genómica de la adaptación. Las bases moleculares tanto de la radiación adaptativa del género como de la especialización trófica son completamente desconocidas. La obtención del genoma de una o varias de estas arañas y su posterior análisis comparativo (entre especies del género, con diferentes especificidades de dieta, y/o con otros arácnidos o artrópodos) nos puede proporcionar información relevante sobre los genes y las vías metabólicas y moleculares que permiten esta diversidad y especialización trófica.

Además, el disponer de la composición, estructura y característica del genoma de arañas del género *Dysdera* puede proporcionar el conocimiento sobre el proceso de la radiación adaptativa y la diversificación del género en el archipiélago canario. Las arañas son organismos no modelo e infrarrepresentados en las bases de datos [138], por lo que la disponibilidad de un nuevo genoma de araña favorecería tanto los análisis evolutivos o comparativos entre artrópodos como el estudio de los eventos de terrestralización [144], la evolución del sistema quimiosensorial en artrópodos [93, 145] como los específicos de arácnidos, especialmente los análisis relacionados con características como el veneno o la seda [140, 141] (Box 7) o para comprender sus relaciones filogenéticas tan polémicas [134–136, 168, 169]. Para esto puede ser muy útil tanto disponer de un genoma completo y bien anotado como de su uso como recurso para obtener marcadores moleculares representativos del género.

1 Introducción

Capítulo 2

Objetivos

A pesar de que los quelicerados son uno de los grandes grupos de artrópodos, son también uno de los grupos menos representados a nivel genómico en las diferentes bases de datos [138]. Los quelicerados, y las arañas en particular, presentan interesantes características biológicas como puede ser la presencia de seda y veneno, la evolución del sistema quimiosensorial y las adaptaciones ocurridas tras el proceso de la terrestralización [93, 139–141, 144, 145]. La disponibilidad de un nuevo genoma de araña favorecerá tanto los análisis evolutivos o comparativos entre artrópodos como en quelicerados o arañas. Además, la disponibilidad de un genoma completo es un recurso importante como referencia para obtener marcadores moleculares representativos del género y poder realizar estudios filogenéticos y de genética de poblaciones.

El principal objetivo de esta tesis doctoral es el de generar recursos, herramientas bioinformáticas e información de metodologías de secuenciación masiva en el estudio de organismos no modelo, en concreto, en arácnidos. En particular, se ha abordado: (i) el desarrollo e implementación de una herramienta bioinformática que facilite la búsqueda de marcadores moleculares en organismos no modelo; (ii) la obtención de la secuencia completa (*de novo*) y su anotación funcional de un genoma de referencia del género *Dysdera*.

2

Los objetivos específicos de esta tesis doctoral son:

- Implementar métodos bioinformáticos para el desarrollo y búsqueda de marcadores moleculares a partir de datos de secuenciación masiva en organismos no modelo.

- Validar mediante simulaciones computacionales exhaustivas, y un conjunto de datos empíricos, los métodos desarrollados.

- Desarrollar una interfaz gráfica para facilitar a los investigadores la búsqueda de marcadores moleculares con las características apropiadas específicas de cada estudio.

- Generar un ensamblaje genómico de alta calidad y continuidad para poder ser utilizado como un representante del género *Dysdera*.

- Determinar la calidad, continuidad e integridad del ensamblaje.

- Anotar estructural y funcionalmente, utilizando bases de datos conocidas y evidencias de ARN, las regiones codificantes de este ensamblaje.

Capítulo 3

Informe de los directores de tesis

3

UNIVERSITAT DE BARCELONA

Dr. Julio Rozas
Catedràtic de Genètica

Departament de Genètica, Microbiologia i Estadística
Facultat de Biologia

Diagonal 643
Edifici Prevosti
08028 Barcelona
Spain

Tel. +34 93 4021495
Fax. +34 93 4034420
jrozas@ub.edu
www.ub.edu/molevol/julio

Informe signat del director de tesi del factor d'impacte dels articles publicats. En cas que es presenti algun treball en coautoria, caldrà incloure també un informe del director de la tesi signat, en què s'especifiqui exhaustivament quina ha estat la participació del doctorant/a en cada article, i si algun dels coautors d'algun dels treballs presentats en la tesi doctoral ha utilitzat, implícitament o explícitament aquests treballs per a la l'elaboració de la tesi doctoral

---

El Drs. **Julio Rozas i Alejandro Sánchez-Gracia,** directors de la Tesi Doctoral elaborada pel Sr. José F. Sánchez-Herrero, amb el títol "**Desarrollo y utilización de herramientas bioinformáticas en el estudio de datos de secuenciación masiva: Análisis genómicos en arácnidos**"

INFORMEN

Que la tesi doctoral està elaborada com a compendi de 4 publicacions amb dades originals (publicacions 1-2 en el cos central de la tesi), i 2 més (publicacions 3-4) a l'apèndix:

Publicacions:

1. Frías-López, C.*, Sánchez-Herrero, J. F.*, Guirao-Rico, S., Mora, E., Arnedo, M. A., Sánchez-Gracia, A. and Rozas, J. 2016. DOMINO: Development of informative molecular markers for phylogenetic and genome-wide population genetic studies in non-model organisms. *Bioinformatics* **32:** 3753-3759.
   Factor d'impacte (5 Year Impact Factor): **[IF = 8.044; Q1 i D1].** Ocupa la posició **2** (sobre **57**) dins la categoria de Mathematical and Computational Biology.
   *, la mateixa contribució

2. Sánchez-Herrero, J. F., Frías-López, C., Escuer, P., Hinojosa-Alvarez, S., Arnedo, M. A., Sánchez-Gracia, A. and Rozas, J. 2019. The draft genome sequence of the spider *Dysdera silvatica* (Araneae, Dysderidae): A valuable resource for functional and evolutionary genomic studies in chelicerates. *GigaScience* **8:** 1-9. Doi: 10.1093/gigascience/giz099.
   Factor d'impacte (5 Year Impact Factor): **[IF = 7.441; Q1].** Ocupa la posició **13** (sobre **69**) dins la categoria de Multidisciplinary Sciences.

3. Romero-Hernández, B., Tedim, A. P., Sánchez-Herrero, J. F., Librado, P., Rozas, J., Muñoz, G., Baquero, F., Cantón, R., Campo, R. D. 2015. *Streptococcus gallolyticus* subsp. *gallolyticus* from Human and Animal Origins: Genetic Diversity, Antimicrobial Susceptibility, and Characterization of a Vancomycin-Resistant Calf Isolate Carrying a *VanA*-Tn*1546*-Like Element. *Antimicrob. Agents Chemother*. **59:** 2006-2015
   Factor d'impacte (5 Year Impact Factor): **[IF = 4.547; Q1].** Ocupa la posició **22** (sobre **123**) dins la categoria de Microbiology.

UNIVERSITAT DE BARCELONA

Dr. Julio Rozas
Catedràtic de Genètica

Departament de Genètica,
Microbiologia i Estadística
Facultat de Biologia

Diagonal 643
Edifici Prevosti
08028 Barcelona
Spain

Tel. +34 93 4021495
Fax. +34 93 4034420
jrozas@ub.edu
www.ub.edu/molevol/julio

4. Bernabeu, M., Sánchez-Herrero, J.F.; Huedo, P.; Prieto, A.; Hüttener, M.; Rozas, J.; Juárez, A. 2019. Gene duplications in the *E. coli* genome: common themes among pathotypes. *BMC Genomics* **20:** 313. doi: 10.1186/s12864-019-5683-4 Factor d'impacte (5 Year Impact Factor): **[IF = 4.142]** (Dades del 2018). Ocupa la posició **58** (sobre **174**) dins la categoria de Genetics & Heredity.

A la publicació 1 (full article a la revista *Bioinformatics*), que també es presenta com a part de la tesi doctoral de Cristina Frías-López, el estudiant de doctorat Jose F. Sánchez Herrero va realitzar una part dels scripts que formen part de la primera versió del software DOMINO, va desenvolupar la GUI, va realitzar la validació computacional (amb simulacions per ordinador), i va contribuir a la redacció del primer esborrany del manuscrit. A la publicació 2, va realitzar la part més important del treball (feines computacionals, analítiques i redacció del primer esborrany del manuscrit). Les publicacions 3-4 (apèndix), han resultat de col·laboracions científiques on el doctorant, fent servir eines computacionals o analítiques desenvolupats en la seva tesi doctoral, ha realitzat una part dels anàlisis bioinformàtics.

Dr. Julio Rozas Liras
Catedràtic de Genètica
Universitat de Barcelona

Alejandro Sánchez-Gracia
Professor Associat de Genètica
Universitat de Barcelona

Capítulo 4

Artículos

4

4.1. DOMINO: development of informative molecular markers for phylogenetic and genome-wide population genetic studies in non-model organisms

El desarrollo de marcadores moleculares es uno de los desafíos más importantes en filogénetica y genética de poblaciones, especialmente en estudios de organismos no modelo. Una estrategia prometedora para la obtención de marcadores apropiados es la utilización de estrategias de partición genómica para la identificación conjunta de un gran número de ellos. Desafortunadamente, no todos los marcadores obtenidos mediante esta estrategia proporcionan suficiente información para resolver muchas de las cuestiones evolutivas a una resolución taxonómica razonable.

Hemos desarrollado la aplicación DOMINO: *"Development of Molecular Markers in Non-Model Organisms"*, una herramienta bioinformática para la obtención de marcadores moleculares, tanto de datos de secuenciación masiva, como de alineamientos pre-computados. Esta aplicación incluye potentes herramientas bioinformáticas ya descritas junto con nuevas implementaciones. Con ello, hemos desarrollado específicamente una *"pipeline"* versátil para descubrir o seleccionar marcadores personalizados a diferentes niveles de resolución taxonómica. Estos marcadores pueden ser utilizados directamente para estudiar los taxones empleados, para amplificaciones posteriores por PCR en estudios que incluyan otros taxones relacionados, o aprovechados como referencia en el diseño de métodos de secuenciación basados en el enriquecimiento del ADN por captura. Evaluamos exhaustivamente el rendimiento de DOMINO y su utilidad al desarrollar marcadores informativos mediante simulaciones computacionales y datos empíricos.

4

Bioinformatics, 32(24), 2016, 3753–3759
doi: 10.1093/bioinformatics/btw534
Advance Access Publication Date: 24 August 2016
Original Paper

Phylogenetics

DOMINO: development of informative molecular markers for phylogenetic and genome-wide population genetic studies in non-model organisms

Cristina Frías-López[1,2,‡], José F. Sánchez-Herrero[1,‡], Sara Guirao-Rico[1,†], Elisa Mora[2], Miquel A. Arnedo[2], Alejandro Sánchez-Gracia[1,*,§] and Julio Rozas[1,*,§]

[1]Departament de Genètica, Microbiologia i Estadística, and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain and [2]Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals, and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona 08028, Spain

*To whom correspondence should be addressed.
[†]Present address: Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Barcelona, Spain

[‡]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

§The authors wish it to be known that, in their opinion, the last two authors should be regarded as joint Last Authors.

Associate Editor: Alfonso Valencia

Received on May 11, 2016; revised on July 7, 2016; accepted on August 9, 2016

Abstract

Motivation: The development of molecular markers is one of the most important challenges in phylogenetic and genome wide population genetics studies, especially in studies with non-model organisms. A highly promising approach for obtaining suitable markers is the utilization of genomic partitioning strategies for the simultaneous discovery and genotyping of a large number of markers. Unfortunately, not all markers obtained from these strategies provide enough information for solving multiple evolutionary questions at a reasonable taxonomic resolution.
Results: We have developed Development Of Molecular markers In Non-model Organisms (DOMINO), a bioinformatics tool for informative marker development from both next generation sequencing (NGS) data and pre-computed sequence alignments. The application implements popular NGS tools with new utilities in a highly versatile pipeline specifically designed to discover or select personalized markers at different levels of taxonomic resolution. These markers can be directly used to study the taxa surveyed for their design, utilized for further downstream PCR amplification in a broader set taxonomic scope, or exploited as suitable templates to bait design for target DNA enrichment techniques. We conducted an exhaustive evaluation of the performance of DOMINO via computer simulations and illustrate its utility to find informative markers in an empirical dataset.
Availability and Implementation: DOMINO is freely available from www.ub.edu/softevol/domino.
Contact: elsanchez@ub.edu or jrozas@ub.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

It is well known that phylogenetic inferences based on a single or very few genetic markers can lead to systematic errors and reach invalid conclusions (Brito and Edwards, 2009; Maddison *et al.*, 1997). Next generation sequencing (NGS) has become a feasible and cost-effective way of obtaining large amounts of genetic markers suitable for addressing ecological and evolutionary questions. Among current methodologies, the hybrid enrichment and the reduction representation sequencing methods (for a review see Lemmon and Lemmon, 2013) are particularly promising approaches for studies in non-model organisms. Markers developed with these methodologies, however, may not be informative enough to resolve multiple evolutionary questions across a reasonable taxonomic range; indeed, some markers may be inefficient for a particular study in a specific taxonomic group, or can be useful only for limited phylogenetic ranges. These problems make often necessary to accomplish various cost-intensive enrichment or reduction representation experiments to further obtain makers suitable to be applicable across a wide range of species.

Recently, some optimizing approaches have been developed to try to overcome this limited marker informativeness. For instance, the *MarkerMiner* 1.0 pipeline (Chamala *et al.*, 2015), outputs different types of multiple sequence alignments (MSA) files, some of them including reference coding sequences containing introns, which facilitates the downstream evaluation of the phylogenetic utility of each marker or the prediction of intron–exon boundaries and intron sizes, very useful for primer or probe of development. Nevertheless, the pipeline does not perform these assessments by itself and the application is specifically devised to work only with transcriptome assemblies and with a set of plant reference genomes. Indeed, the possibility of selecting particular markers with a specific number of samples has been recently implemented in the RAD-Seq data processing pipeline *RADIS* (Cruaud *et al.*). However, this application does not include other key options and parameter combinations, such as the selection of a specific nucleotide variation range across a set of pre-defined taxa, options that can be very useful for a plethora of studies. *BaitFisher* (Mayer *et al.*, 2016) also implements a novel approach to optimize the design of target enrichment baits to be applicable across a wide range of taxa. This software includes an algorithm to infer target DNA enrichment baits from multiple taxa by exploiting user-provided nucleotide sequence information of target loci in a representative set of species and can handle both genomic and cDNA data. Nevertheless, this software works on the basis of MSA of already known target loci that directly serves as templates for bait design (i.e. it cannot be used with raw NGS data or for *de novo* marker discovery).

Here we present Development Of Molecular markers In Non-model Organisms (DOMINO) a new bioinformatics tool that facilitates the development of highly informative markers from different data sources, including raw NGS reads and pre-computed MSA in various formats (such as those from RAD data). DOMINO efficiently process NGS data or pre-computed MSA and identifies (i.e. *de novo* discovery) or selects the sequence regions or alignments that meet user-defined criteria. Customizable features include the length of variable and conserved regions (when requested), the minimum levels (or a preferred range) of nucleotide variation, how to manage polymorphic variants, or which taxa (or what fraction of them) should be covered by the marker. All these criteria can be easily defined in a user-friendly graphical user interface (GUI) or under a command-line version that implements some extended options and that it is particularly useful for working with large NGS datasets in high performance computers (Supplementary Fig. S2; see also the DOMINO documentation). The regions identified or selected in DOMINO can be (i) directly used as markers with a particular depth of taxonomic resolution, (ii) utilized for their downstream PCR amplification in a broader taxonomic scope or (iii) used as suitable templates to optimized bait design for target DNA enrichment techniques.

2 Methods and implementation

2.1 DOMINO workflow

The DOMINO workflow consists of four main phases (Fig. 1) that can be run either using the DOMINO GUI or the extended command-line version (see the DOMINO manual in the DOMINO Web page). In both cases, the most relevant results from each phase are conveniently reported in the appropriate output files.

2.1.1 Input data and pre-processing phase

DOMINO accepts input sequence data files in two different formats, the 454 Pyrosequencing Standard Flowgram Format (SFF), and FASTQ format (Cock *et al.*, 2010). These input files can contain 454 or Illumina (single- or paired-end) raw reads from *m* taxa (the 'taxa panel'). The sequences from each taxon should be properly identified with a specific barcode (aka, tag, MID or index), or loaded in separate files, also appropriately named (see the DOMINO manual in the DOMINO Web site for details). DOMINO is designed to filter low quality, low complexity, contaminant and very short reads using either default or user-specified filtering parameters. Mothur, PRINSEQ, NGS QC toolkit, BLAST, as well as new Perl functions specifically written for DOMINO (DM scripts) are used to perform these tasks (Supplementary Table S1). DOMINO uses Mothur v1.32.0 (Schloss *et al.*, 2009) to extract reads from SFF files and store them in FASTQ format, which are subsequently converted to FASTA and QUAL files. Low quality or very short reads are trimmed, or definitely removed, using NGS QC Toolkit v2.3.1 (Patel and Jain, 2012). PRINSEQ v0.20.3 package (Schmieder and Edwards, 2011) is used to eliminate low complexity reads using the implemented DUST algorithm. Putative contaminant sequences, such as bacterial DNA frequently found in genomic samples (Leese *et al.* 2012), cloning vectors, adaptors, linkers and high-throughput library preparation primers, can also be removed using a DOMINO function that performs a BLAST search (BLAST v2.2.28) (Altschul *et al.*, 1990) against UniVec database (http://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/) and/or against a user-supplied contaminant database (see the DOMINO manual).

2.1.2 Assembly phase

When working with just NGS reads, the program first applies an assembly-based approach; the pipeline is therefore optimized to work with genome partitioning methods in which the length of the size-selected (or enriched) fragments and the sequencing depth are enough to permit the assembly of a set of homologous fragments. For data from restriction-site associated DNA (RAD) sequencing and related methods see the Mapping/Alignment phase section. DOMINO performs separate assemblies, one for each panel taxon, using MIRA v4.0.2 (Chevreux, 1999), either with the pre-processed reads from the previous step or with those supplied by the user. Although the default parameter values vary in function of the particular sequencing technology, the majority of them are shared (see the DOMINO manual). In order to avoid including repetitive and chimeric regions, all contigs (and the corresponding reads) identified as repeats in the MIRA algorithm are discarded from the mapping/

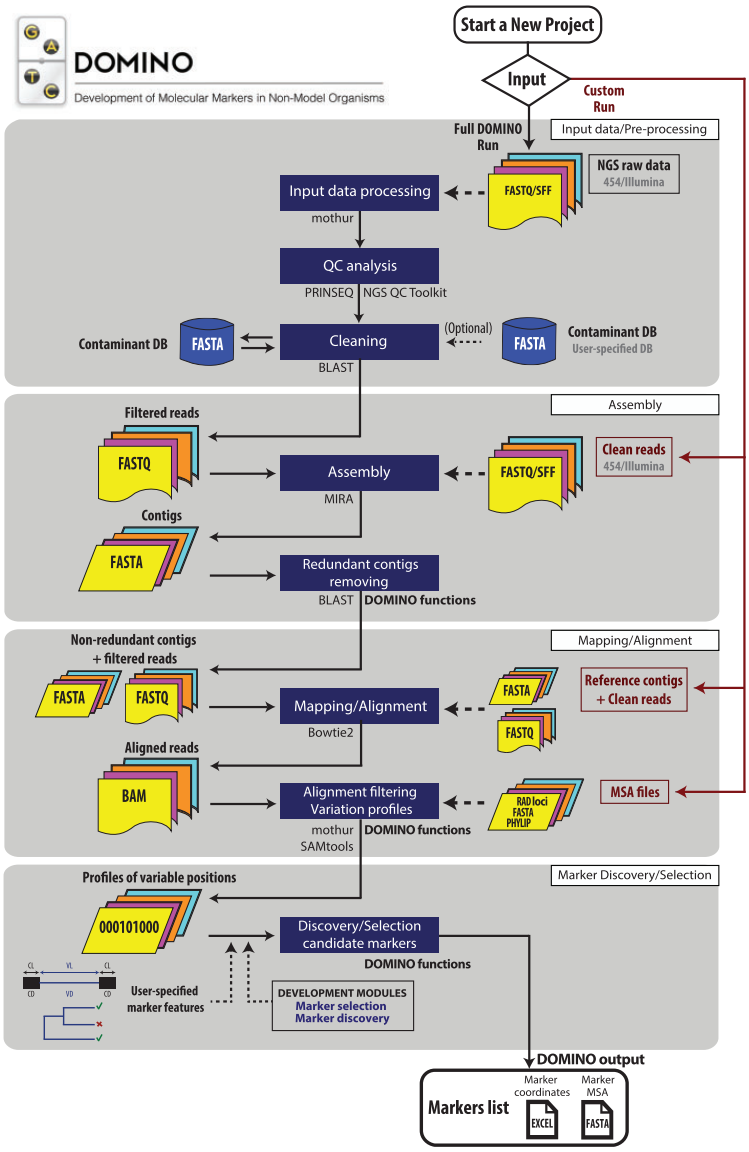


Fig. 1. Workflow showing the basic steps used to discover or select molecular markers with the DOMINO software

alignment phase (Chevreux, 1999). Since MIRA can generate redundant contigs because of polymorphic and paralogous regions, we have implemented a specific DOMINO function that performs a clustering of all contigs based on an all versus all contigs BLAST search to identify and remove such redundancies. The DOMINO command line version (see below) also includes an option to perform a second iterative assembly step using the software CAP3 (Huang, 1999). If selected, this option uses MIRA output sequences (contigs and singletons) as input for CAP3 under a relaxed parameter scheme.

2.1.3 Mapping/alignment phase

DOMINO uses *Bowtie2* (Langmead and Salzberg, 2012) to map the pre-processed reads from each taxon to the assembled contigs of the other $m-1$ taxa from the panel. Thus, in this step, DOMINO builds $m(m-1)$ sequence alignment/map files (SAM/BAM files. In the case of a panel of $m=4$ taxa, e.g. DOMINO will build $4 \times 3 = 12$ SAM/BAM files during this step. The reason behind this particular mapping strategy lies in the dissimilar performance of alignment/mapping algorithms depending on the divergence between the reads and the reference sequences. Immediately after generating BAM files, DOMINO

Fig. 2. Screenshot of Marker Discovery/Selection TAB included in the DOMINO GUI

removes all unmapped contigs and multi-mapping reads. This step is critical to avoid alignment artifacts, which can create false positive markers (i.e. sequence regions with misleading high levels of nucleotide diversity). The contigs with an unusually large number of aligned reads, which can correspond to repetitive regions, are also removed (they are not suitable for designing single copy markers) Later, DOMINO will build one pileup file per each BAM/SAM file using the SAMtools v0.1.19 suite (*mpileup* option) (Li *et al.*, 2009).

Since sequencing errors might have a great effect on the marker selection, DOMINO incorporates their own functions for detecting and masking putative sequencing errors, which apply a very conservative criterion for variant calling. First, to avoid the calling of spurious nucleotide variants in low sequencing coverage experiments (i.e. erroneously assigned variants fixed between the taxa from the panel), DOMINO mask the information from positions with only one read mapped to the reference. Furthermore, sequencing errors may also inflate the number of called polymorphisms under the Polymorphic Variants option in the marker identification/selection phase. To avoid such undesirable effect, DOMINO incorporates a similar conservative criterion to use only highly credible polymorphisms. Under the Polymorphic Variant option, DOMINO will assume that each taxon represents a diploid individual; for positions with eight or more reads mapped, DOMINO discards those polymorphic variants in which the frequency of the minor allele is significantly lower than the expected under error free data (hence, in absence of sequencing errors the distribution of observed nucleotide counts at each position would follow a binomial distribution). For lower coverage values, DOMINO will use the information of a polymorphic variant only if the allele with the minor frequency is present in two or more reads. This testing procedure, applied independently for each position within each species, will likely discard some true polymorphic sites; this variant calling approach, however, makes DOMINO highly conservative in detecting true markers when including polymorphisms in the analysis (i.e. DOMINO will use only highly confident within-species segregating variants for the marker Discovery/Selection phase). Ambiguity codes, either introduced by MIRA assembler in contig sequences or present in user-supplied reference sequences or MSA, are also considered by DOMINO to decide whether a position is or not variable.

After applying all the above-mentioned post-mapping filters, DOMINO combines the variation profiles (arrays with the information about the state of each position, conserved or variable between taxa pairs) obtained from each of the $m-1$ pileup files including the same reference sequence (i.e. the same taxon), into a single multiple taxa variation profile (MTVP). Since each of these references will be likely fragmented in i contigs, DOMINO will build $i \times m$ MTVP per taxon. Each of these MTVP will be independently scanned for regions containing candidate markers in the next phase. If the user provides reference sequences from a single taxon (e.g. a genome draft), plus the reads from the m different taxa, the program builds only one MTVP set (one per contig or scaffold in the supplied reference). On the other hand, if the input includes a single or multiple pre-computed MSA instead of NGS data, DOMINO skips the alignment/mapping phase and directly generates the single MTVP set (one per aligned region). In this point, the program accepts MSA files in FASTA (multiple FASTA files, one per linked region), PHYLIP (multiple PHYLIP files, one per linked region, or one multi PHYLIP file with the alignment of all regions) and pyRAD LOCI (*.loci files generated by the program pyRAD; Eaton, 2014) and STACKS fasta (batch_X.fa output files generated from the population analyses in the program STACKS; Catchen *et al.*, 2011) output files.

2.1.4 Marker discovery/selection phase

Each MTVP generated in the previous step is either scanned for the presence of candidate marker regions using a sliding window approach (DOMINO marker discovery module), or used to select markers (with the desired features) among the MSA loaded in the previous tab (DOMINO marker selection module). In the first case, a specific DOMINO function searches for sequence regions of desired length (Variable region Length, VL), showing the minimum level of variation indicated by the user (Variable region Divergence, VD). DOMINO can also restrict that this variable region was flanked (or not) by highly conserved regions (Conserved region Divergence, CD) of a predefined length (Conserved region Length, CL); an information useful to further design PCR primers. Moreover, DOMINO can strictly restrict the search to a particular set of taxa (from the panel), or just specify the minimum number of taxa required to be covered by the marker (by changing the Minimum number of Covering Taxa parameter; MCT $< m$). As indicated, DOMINO can use or not the information from polymorphic sites. An appropriated combination of selected taxa and MCT and VD parameter values will allow the user select a large set of informative markers suitable to be applicable across a wide range of taxa. In the second case, the DOMINO selection module allows directly selecting the most informative markers among the loaded by the user in the same way and with the same personalized features described above. For RAD loci, a particular range of variable positions (VP) between the closest taxa (instead of the VD parameter) must be specified. This option allows selecting informative RAD loci while excluding those exhibiting anomalous high levels of variation, which might reflect RAD tag clustering errors. The specific selection of a set of loci/MSA that meet some specific phylogenetic criteria using the DOMINO selection module can be very helpful to further design probes for different target enrichment techniques, including the enrichment of specific RAD segments using hyRAD (Suchan *et al.*, 2016).

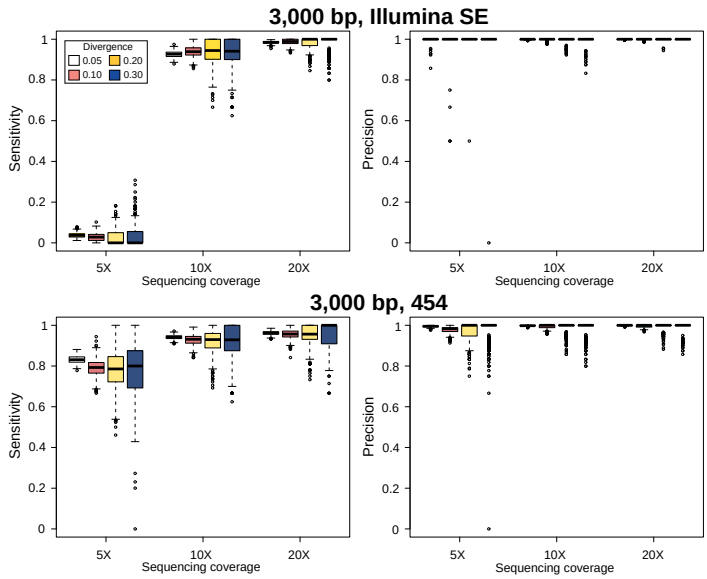


Fig. 3. Sensitivity and precision estimates for simulated datasets of 100 fragments of 3 kb after their *in silico* sequencing with Illumina and Roche-454 technologies

After the last phase, DOMINO reports the list the genomic regions (and their coordinates) or MSA that meet the selection criteria, along with the corresponding MSA of these regions for the selected taxa. Since DOMINO can work with more than one MTVP set (*m* in a full DOMINO run), some of the markers found in MTVP based on different reference taxa may be redundant (they can cover the same genomic region, although with different coordinates; see Mapping/Alignment phase section), while other can be found only in one particular profile. To avoid reporting redundant information, we have implemented a BLAST-based function to collapse these maker sequences, only reporting unique markers. To maximize the probability of finding informative markers, the final list of candidates under the DOMINO marker discovery module can include overlapped regions that fulfill the specified characteristics. Operationally, all regions that meet the criteria for being considered a candidate marker (after moving the scanning window five or more base pairs) are listed as different markers in the final output. In this way, users can choose the best marker to be used directly for further analyses or the more appropriated region of each contig to be PCR amplified and sequenced in additional focal species (i.e. the best marker from each linked block).

2.2 DOMINO GUI

DOMINO can be run either in the command prompt, by setting a large set of command line options, or using the GUI specifically developed to facilitate its use to non-experts in NGS bioinformatics tools (Fig. 2; see also the DOMINO manual for details). The DOMINO GUI is a cross-platform application that allows the user to interactively set marker selection criteria by tuning the most important parameters and options available in the command prompt version. It should be noted that for huge NGS datasets (which require substantial amounts of computational resources) a full DOMINO run using the GUI version is not recommendable. In this case, the user can either run DOMINO under the command line version using high performance computer clusters or, take advantage of the custom run options available in the GUI version to enter in DOMINO partially processed data, e.g. pre-processed reads, assemblies or alignment files (SAM/BAM) obtained with other memory-efficient software (Supplementary Table S2).

2.3 System and availability

The GUI was built using the cross-platform library and user interface framework Qt (https://www.qt-project.org/) based on C++ scripting language. Since most of the functions specifically developed for this work are implemented in Perl scripting language, users need to install first a recent version of Perl (version 5.12 or higher; http://www.perl.org/). The source code, the documentation and some example data files are freely distributed under the GNU GPL software license at: http://www.ub.edu/softevol/domino.

3 Results and conclusions

3.1 Computer simulations

We conducted an exhaustive computer simulation study to assess the performance of DOMINO in detecting informative markers (i.e. simulated regions that meet specific marker selection criteria) from NGS data. For that, we emulated an RRL-like experiment of four closely related species exhibiting different levels of nucleotide divergence among them and incorporating substitution rate heterogeneity across sites to create genuine informative markers. The topology of the species tree used for the simulations was fixed (Supplementary Fig. S1). In each replicate, we generated an independent RRL-like dataset of 100 fragments, of different length (3 or 10 kb) each. The nucleotide sequences were simulated with the program evolver, included in the PAML v4.7 package (Yang, 1997, 2007), using 0.1, 0.15, 0.20 or 0.30 substitutions per site between the two most

divergent sequences, under the Jukes and Cantor (1969) substitution model with substitution rate heterogeneity across sites (modeled as a discrete gamma with 10 categories and $\alpha = 0.01$). For each replicate, we simulated a complete NGS experiment in the Roche-454 (reads with an average length of ~400 bp), and the Illumina HiSeq2000 platforms (average length of 101 bp; single and paired-ends) using the ART v2.5.8 program (Huang *et al.*, 2012) with default parameters and three different sequencing coverage values (5×, 10× and 20×). We generated 500 simulation replicates for each of the 48 possible scenarios (i.e. for each combination RRLs fragment length, divergence, sequencing platform and coverage values), resulting in a total 27 000 DOMINO runs, which took roughly 80 000 CPU h.

Using the DOMINO marker discovery module under the command line version, we first traced the number and the location of the regions that meet the selection criteria present in each simulated fragment previous to emulate their NGS sequencing (true markers; TNM). Subsequently, for each dataset, we execute a full run of our program using the simulated NGS reads to obtain the list of candidate markers (detected markers; DNM) for each scenario. For this experiment, we define an informative marker as a variable region of 600 bp (VL = 600), present in all four species (MCT = 4), showing at least 0.01 nucleotide substitutions per site between any pair of species (VD = 0.01), and flanked by two highly conserved regions of 60 or more bp long (CL = 60; only one substitution across species was permitted; CD = 1). We assessed the performance of DOMINO in detecting the TNM by measuring the sensitivity and precision in each replicate and plotting their distribution across the 500 replicates (Fig. 3; Supplementary Material).

We found that DOMINO pipeline has a high sensitivity in detecting the existing TNM, yielding averages of true positive rates values >0.9 for Illumina reads and when coverage values are equal or higher than 10× (Fig. 3). As expected, lower coverage values (5×) result in a reduction of the sensitivity estimates; in this case, DOMINO runs using 454 long reads outperforms those using Illumina short reads (e.g. average sensitivities close to 0.8 for the 454 under all tested nucleotide divergences in 3 kb fragments). Noticeably, we found that DOMINO show high sensitivities even for relative high divergence levels (up to 0.3 substitutions per site between the two more diverged taxa); in this case, the program performs slightly better when using short reads as input. In the light of this high sensitivity, precision becomes a critical aspect to be considered for further successful marker discovery. We found that DOMINO also detects TNM regions with high precision (most values are close to 1 regardless of the condition), yielding very few number of false positives. The performance of DOMINO when using reads from larger library fragments (10 kb) is very similar to that of the observed for 3 kb (Supplementary Fig. S2).

3.2 Application to empirical data

To illustrate the utility of DOMINO on real biological data, we performed a RRL sequencing experiment (using 454 reads; see Supplementary Material for details), which allow running all phases of the application and the DOMINO marker identification module, from raw reads to marker selection. We used four individuals (panel with four taxa) belonging to the spider family *Nemesiidae* (Araneae) for this analysis (Supplementary Material, Fig. S3). We identified many candidate regions that fulfill the requested marker characteristics (Supplementary Tables S3–S6), and tested the suitability of six of them by PCR amplification and Sanger sequencing in a larger panel that also included other 14 phylogenetically related species (focal species). The obtained phylogenetic tree not only recovered the expected relationships among the taxa from the panel but also demonstrates that the sequenced markers are useful to establish the phylogenetic relationships of the focal ones (Supplementary Fig. S4).

3.3 Conclusions

DOMINO will assist researches working with non-model organisms in the development of molecular markers for DNA variation studies. First, it allows obtaining a list of 'personalized' markers that meet user specific criteria without the mandatory need of a reference genome, which will improve their application from highly specific taxonomic scopes to more wide phylogenetic ranges. Second, its output alignment files, jointly with the information about markers coordinates and features provided by the program, can be either directly utilized in variation studies, or used as a templates for further downstream PCR amplification or target DNA enrichment probe design. Third, the DOMINO GUI makes this application accessible and easy-to-use to non-experts in the bioinformatics of NGS data handling and analysis. Finally, DOMINO is open cross-platform software that can be straightforwardly adapted to other pipelines or used in high performance computers. Although current version of the program works with raw reads of a limited number of reduction representation schemes (e.g. DOMINO cannot process raw reads from RAD- or RNA-Seq approaches) and sequencing platforms (Illumina short and 454 long reads), the modular structure of DOMINO will allow easily expanding the software to accept NGS data from other sources.

Funding

This work was supported by the Ministerio de Educación y Ciencia of Spain (No. BFU2010-15484 and No. CGL2013-45211 to J.R. and No. CGL2012-36863 to M.A.A.).

References

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Brito,P.H. and Edwards,S.V. (2009) Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica*, **135**, 439–455.

Catchen,J.M. *et al.* (2011) Stacks: building and genotyping Loci de novo from short-read sequences. *G3 (Bethesda)*, **1**, 171–182.

Chamala,S. *et al.* (2015) MarkerMiner 1.0: A new application for phylogenetic marker development using angiosperm transcriptomes. *Appl. Plant Sci.*, **3**, 1400115.

Chevreux,B. *et al.* (1999) Genome sequence assembly using trace signals and additional sequence information. *Comput. Sci. Biol. Proc. German Conf. Bioinform.*, **99**, 45–56.

Cock,P.J.A. *et al.* (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.*, **38**, 1767–1771.

Cruaud,A. *et al.* (2016) *RADIS*: analysis of *RAD*-seq data for interspecific phylogeny. *Bioinformatics*, doi:10.1093/bioinformatics/btw352.

Eaton,D.A.R. (2014) PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics*, **30**, 1844–1849.

Huang,W. *et al.* (2012) ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**, 593–594.

Huang,X. (1999) CAP3: a DNA sequence assembly program. *Genome Res.*, **9**, 868–877.

Jukes,T.H. and Cantor,C.R. (1969). Evolution of protein molecules. In: Munro, H.N. (ed.) *Mammalian Protein Metabolism*. Academic Press, New York, pp. 21–132.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

Leese,F. *et al.* (2012) Exploring Pandora's box: potential and pitfalls of low coverage genome surveys for evolutionary biology. *PLoS One*, **7**, e49202.

Lemmon,E.M. and Lemmon,A.R. (2013) High-throughput genomic data in systematics and phylogenetics. *Annu. Rev. Ecol. Evol. Syst.*, **44**, 99–121.

Li,H. *et al*. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Maddison,W.P. *et al*. (1997) Gene trees in species trees. *Syst. Biol*., **46**, 523–536.

Mayer,C. *et al*. (2016) BaitFisher: a software package for multispecies target DNA enrichment probe Design. *Mol. Biol. Evol*., **33**, 1875–1886.

Patel,R.K. and Jain,M. (2012) NGS QC toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One*, **7**, e30619.

Schloss,P.D. *et al*. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol*., **75**, 7537–7541.

Schmieder,R. and Edwards,R. (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–864.

Suchan,T. *et al*. (2016) Hybridization Capture Using RAD Probes (hyRAD), a new tool for performing genomic analyses on collection specimens. *PLoS One*, **11**, e0151651.

Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci*., **13**, 555–556.

Yang,Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol*. **24**, 1586–1591.

4

DOMINO: Development of informative molecular markers for phylogenetic and genome-wide population genetic studies in non-model organisms

Supplementary Information

Cristina Frías-López[1,2,†], José F. Sánchez-Herrero[1,†], Sara Guirao-Rico[1,3], Elisa Mora[2], Miquel A. Arnedo[2], Alejandro Sánchez-Gracia[1,‡], and Julio Rozas[1,‡,*]

[1]Departament de Genètica, Microbiologia i Estadística, and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain.[2]Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals, and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain.[3]Current affiliation: Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Barcelona, Spain.

*To whom correspondence should be addressed.

[†]The authors wish it to be known that in their opinion, the first two authors should be regarded as joint First Authors.

[‡]The authors wish it to be known that in their opinion, the last two authors should be regarded as joint Last Authors.

Computer Simulations

We assessed the performance of DOMINO in detecting the genuine makers present in the simulated sequences (i.e., the regions in these sequences that meet the marker selection criteria to be further specified in DOMINO) by measuring the sensitivity and precision in each independent replicate and by plotting their distribution in the 500 replicates (Figure 3; Supplementary Fig. S2), being:

Sensitivity = TP / (TP+FN) and
Precision = TP / (TP+FP),

where, TP (true positives) indicates the number of simulated regions meeting the criteria that are correctly identified, FN (false negatives) indicates the number of these genuine marker regions that were not detected by the DOMINO discovery module after read assembly and mapping and FP (false positives) indicates the number of simulated regions incorrectly identified as markers.

Empirical Data

As an example of the application of DOMINO to empirical NGS data, we used the program to search for highly informative markers suitable to be used in phylogenetic studies in the spider family *Nemesiidae* (Araneae, Mygalomorphae). *Nemesia* and. For that, we chose a taxa panel of four species, consisting in three *Nemesia* (Audouin, 1826) and one *Iberesia* (its putative sister group, *Iberesia* Decae & Cardoso, 2006) samples (Supplementary Fig. S3). Specifically, we included in the panel two individuals from two different populations of *Nemesia raripilia* (Simon, 1914; *Nemesia raripilia* populations 061 and 079; collected in Coll de les Tres Creus, Sant Llorenç del Munt i Serra de l'Obac Natural Park, Barcelona, Spain), one individual of a different unidentified *Nemesia* species (*Nemesia* sp population 043 from Cabrera de Mar - Barcelona, Spain) and one individual of *Iberesia brauni* (L. Koch, 1882) (locality 098, collected in Port de Soller, Majorca, Spain).

We digested the genomic DNA of these four individuals with the eight-cutter restriction enzyme *Not*I (restriction site 5' GC/GGCCGC 3'). Fragments ranging from 2.5 to 3 kb were selected to construct the representation libraries by excising the corresponding bands from

the agarose gel, following by purification with the Qiagen Gel Extraction Kit (Qiagen). The Illustra GenomiPhi V2 Amplification Kit (GE Healthcare) was used to increase the amount of recovered DNA following the manufacturer's specifications. The amplified DNA was treated with RNase (Qiagen), and subsequently purified using the Qiagen PCR Purification Kit (Qiagen). The purified DNA sample was quantified with the Qubit® 2.0 Fluorometer (QBIT Assays, Invitrogen). The sequencing was conducted on a 454/Roche GS-FLX Titanium, with each sample individually tagged using the Roche 454 Pyrosequencing MID (Multiplex Identifier DNA) tags, and using 1/2 picotitre plate. Assuming that the restriction sites are randomly distributed across the genome, we estimated that the libraries represent ~0.007 of each genome (~21 Mbp, assuming a ~3Gbp genome). We obtained ~425,000 reads, and used DOMINO to pre-process this raw data, and to the *de novo* assembly each RRL fragment Supplementary Table S3). We searched, applying different parameter settings, for regions candidate to encompass suitable makers (Supplementary Table S4). We tested the suitability of six of the identified candidate markers by PCR amplification in individuals from the same four species of the panel along with other 14 phylogenetically related species (focal species). After the Sanger sequencing of each fragment, we built an MSA per each marker region using the program Mafft (Katoh *et al.*, 2002; Katoh and Standley, 2013), which were further concatenated to obtain the final MSA used for the phylogenetic analysis in RAxML version 8 (Stamatakis, 2014) (Supplementary Fig. S4).

DNA Deposition

The data have been deposited in the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/), with number: PRJNA327555

References

Katoh,K. et al. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res., **30,** 3059–66.

Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol., **30,** 772–80.

Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, **30,** 1312–3.

Supplementary Figures

Figure S1. Topology and relative branch lengths of the tree used to simulate sequence data. In this example, we show the tree used to simulate sequences with 0.2 nucleotide substitutions per site between the two more distant taxa.

Figure S2. Sensitivity and precision estimates for data sets of 100 fragments of 10 kb after their in silico sequencing of simulated fragments with Illumina and Roche-454 technologies.

Figure S3. Maximum likelihood phylogenetic tree showing the relationships among the four species included in the taxa panel. This tree was built using a multiple sequence alignment of COI sequences. Branch lengths are in nucleotide substitutions per site.

Figure S4. Maximum likelihood phylogenetic tree showing the relationships among the four taxa included in the panel and other 11 focal species. The tree was built using a concatenated multiple sequence alignment with the sequence information of six of the markers identified by DOMINO.

Supplementary tables

Table S1. Summary of software, algorithms and Perl functions included in DOMINO.

Table S2. RAM, CPU and execution times of different DOMINO runs and data sets.

Table S3. Summary statistics of the analysis of the NGS data from the four *Nemesiidae* species

Table S4. Results of the DOMINO maker discovery module using the NGS data from the four *Nemesiidae* species

Table S5. Summary statistics of the analysis of a subset of 4,000 reads (NGS data from the four *Nemesiidae* species).

Table S6. Results of the DOMINO maker discovery module using the subset of 4,000 reads (NGS data from the four *Nemesiidae* species).

SUPPLEMENTARY FIGURES

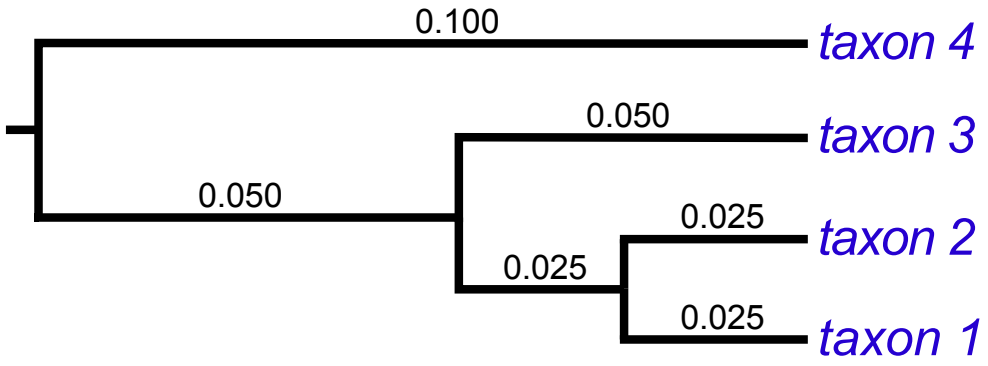


Figura S1

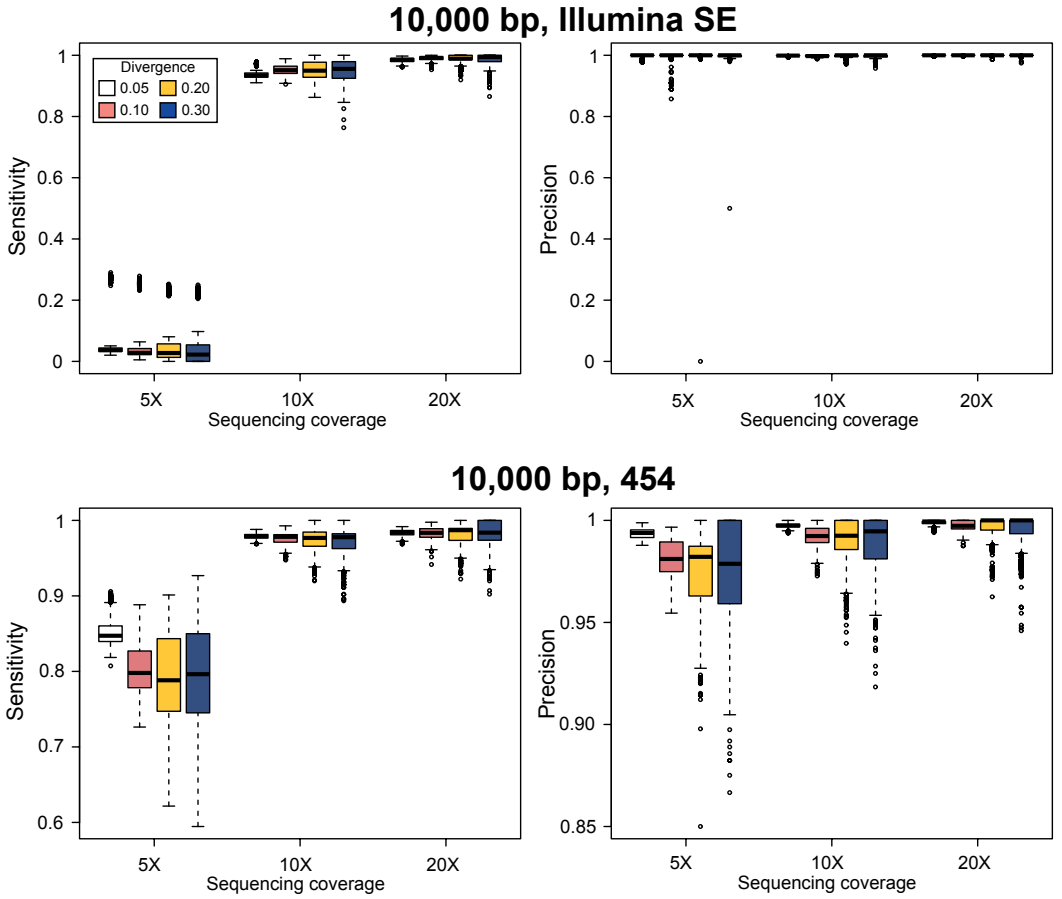


Figura S2

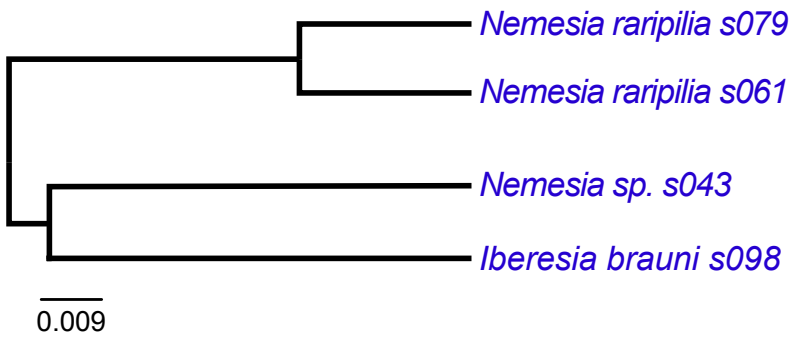


Figura S3

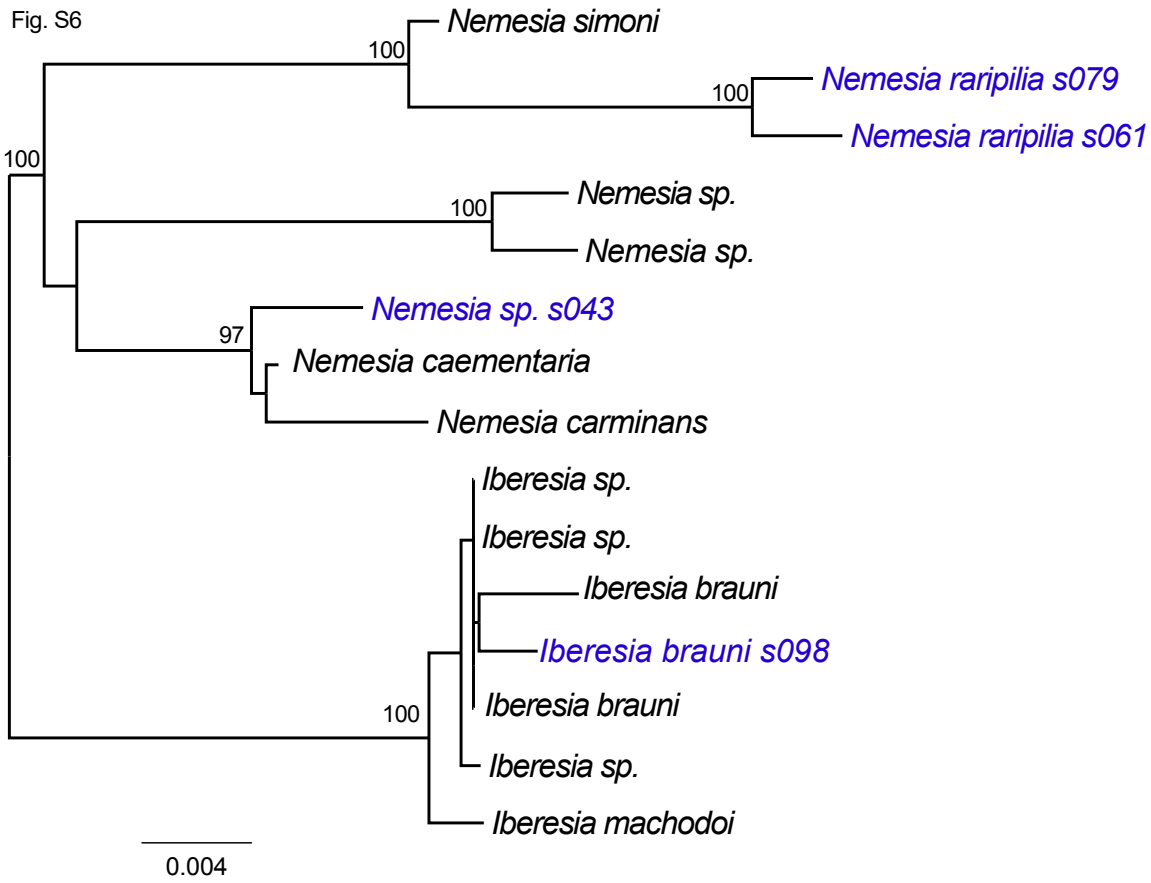


Figura S4

SUPPLEMENTARY TABLES

Table S1. Summary of software, algorithms and Perl functions included in DOMINO

A) Software and algorithms used by DOMINO

	Version	Task	Reference
ART_454	2.5.8	Computer simulation of NGS reads	W. Huang, Li, Myers, & Marth, 2012
BLAST	2.2.28	Contamination search	Altschul, Gish, Miller, Myers, & Lipman, 1990
Bowtie2	2.2.3	Alignment/Mapping	Langmead & Salzberg, 2012
CAP3	Date 21/12/2007	Assembly	X. Huang & Madan, 1999
Evolver (PAML package)	4.7	Computer simulation of MSA	Yang, 2007
MIRA	4.0	Assembly	Chevreux, Wetter, & Suhai, 1999
mothur	1.32.0	SFF data file processing	Schloss et al., 2009
NGS QC Toolkit	2.3.1	Quality Filtering	Patel & Jain, 2012
PRINSEQ	0.20.4	Quality Filtering, graph generation and FASTQ extraction/generation	Schmieder & Edwards, 2011
SAMTools	0.1.19	SAM/BAM data files processing and pileup file generation	Li et al., 2009

B) Perl modules used by DOMINO

	Version
File-Path	2,09
List-MoreUtils	0.410
List-Uniq	0.20
Math-CDF	0,1
PathTools	3,47
Scalar-List-Utils	1,39
Spreadsheet-WriteExcel	2.40
Exporter-Tiny	0,042
GD	2,56
GDTextUtil	0,86
Parallel String-Approx	3,26

Table S2. RAM, CPU and execution times of different DOMINO runs and data sets

Type of run	Taxa panel	Input data[1]	Computational resources #cores (total RAM in GB)[2]	Max. memory (in GB)	Max. CPUs used	Time (minutes)
Full DOMINO run	Taxa panel with four *Nemesiidae* species	425,065 reads (Roche 454)	4(8)	1.8	4	100.21
			8(8)	1.8	5	141.45
Full DOMINO run	Taxa panel with four *Nemesiidae* species	4,000 reads (Roche 454)	4(8)	0.61	1	3.47
			8(8)	0.78	1	4.16
DOMINO selection module[3]	Taxa panel with 24 individuals	27727 RAD loci	4(8)	0.3	1	14.51
			8(8)	0.6	1	12.09

[1]. Data Sets

Roche 454. Data described in the paper

RAD Loci. Data set: pyRAD loci output: c88_d6m10p2_wRE.readloci

Downloaded from Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.ts2hj

Described in Hipp et al. (2014) PLoS ONE 9: e93975.http://dx.doi.org/10.1371/journal.pone.0093975

[2]. Hardware-Operating System

4(8). Linux with Ubuntu 14.04 (64 bits); Processor Intel Core i5 650@3.20 GHz x 4; with 8 GB RAM

8(8). Linux with Debian 7; Processor Intel Xeon E5420@2.5 GHz x 8; with 8 GB RAM

[3]. RAD Options

Selection of RAD loci, having at least 5 individuals (MCT = 5) exhibiting 1-5 variable positions (VP 1::5; i.e., VP min = 1; VP max = 5)

Table S3. Summary statistics of the analysis of the NGS data from the four *Nemesiidae* species

	N_raripilia061	*N_raripilia079*	*Nemesia_sp043*	*I_brauni098*	TOTAL
Raw data (reads)	166.259	88.021	85.331	85.454	425.065
Base pairs	69.440.459	38.443.530	36.654.576	33.043.616	144.538.565
N50	609	607	608	594	605
Pre-processing					
Filtered rea	109.763	60.284	56.194	55.361	281.602
Base pairs	48.733.514	27.555.400	25.258.430	22.957.648	124.504.992
N50	579	576	575	558	572
Assembly (MIRA)					
Assembled	73.879	38.887	34.044	31.787	178.597
Contigs	15.575	8.540	7.583	7.016	38.714
DOMINO post-assembly filtering					
Filtered con	10.003	6.158	5.882	5.400	27.443
Base pairs	5.913.109	3.658.207	3.389.037	2.850.000	15.810.353
N50	694	705	692	656	687

*, Number of reads and contigs after the pre-processing and post-assembly filtering steps

Table S4. Results of the *DOMINO maker discovery module* using the NGS data from the four *Nemesiidae* species

Number of contigs with candidate markers (and the total number of markers, in parenthesis) identified under different settings

	VL = 300	VL = 400	VL = 500	VL = 1000
***n* = 4*, MCT**	1 (1)	2 (2)	0 (0)	0 (0)
***n* = 4*, MCT**	16 (23)	7 (12)	4 (6)	2 (3)
***n* = 3#, MCT**	9 (15)	5 (8)	2 (4)	1 (2)

*, selecting all four focal taxa
#, selecting only three taxa (*N_raripilia061*, *I_brauni098*, *Nemesia_sp043*) from the four focal species

Parameters values fixed
CL = 30
CD = 1
VD = 0.01

Parameters definition
CL, sequence length of the conserved region
CD, maximum number of differences accepted in the conserved region
VL, sequence length of the variable (marker) region
VD, minimum pairwise divergence accepted among the selected sequences
MCT, minimum number of covering taxa per marker

Table S5. Summary statistics of the analysis of a subset of 4000 reads (NGS data from the four *Nemesiidae* species)

	N_raripilia061	*N_raripilia079*	*Nemesia_sp043*	*I_brauni098*	TOTAL
Raw data (reads)	1.000	1.000	1.000	1.000	4.000
Base pairs	538.593	530.296	543.918	460.673	2.073.480
N50	613	613	623	596	611
Pre-processing					
Filtered rea	996	995	994	996	3.981
Base pairs	536.748	528.179	540.660	458.780	2.064.367
N50	613	614	622	596	611
Assembly (MIRA)					
Assembled	974	971	938	924	3.807
Contigs	100	128	134	145	507
DOMINO post-assembly filtering					
Filtered con	57	100	187	170	514
Base pairs	57.266	81.757	147.530	109.829	396.382
N50	1.173	1.100	965	864	1.026

*, Number of reads and contigs after the pre-processing and post-assembly filtering steps

Table S6. Results of the *DOMINO maker discovery module* using the subset of 4000 reads (NGS data from the four *Nemesiidae* species)

Number of contigs with candidate markers (and the total number of markers, in parenthesis) identified under different settings

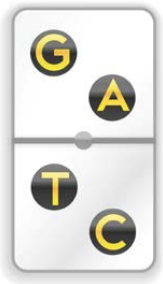
	VL = 300	VL = 400	VL = 500	VL = 1000
***n* = 4*, MCT**	0 (0)	0 (0)	0 (0)	0 (0)
***n* = 4*, MCT**	10 (16)	6 (9)	3 (4)	2 (3)
***n* = 3#, MCT**	6 (10)	5 (7)	3 (5)	0 (0)

*, selecting all four focal taxa
#, selecting only three taxa (*N_raripilia061*, *I_brauni098*, *Nemesia_sp043*) from the four focal species

Parameters values fixed
CL = 30
CD = 1
VD = 0.01

Parameters definition
CL, sequence length of the conserved region
CD, maximum number of differences accepted in the conserved region
VL, sequence length of the variable (marker) region
VD, minimum pairwise divergence accepted among the selected sequences
MCT, minimum number of covering taxa per marker

MANUAL



DOMINO

Development of Molecular Markers in Non-Model Organisms

Cristina Frías-López
José F. Sánchez-Herrero
Miquel A. Arnedo
Alejandro Sánchez-Gracia
Julio Rozas

Departament de Genètica, Microbiologia i Estadística
Departament Biologia Evolutiva, Ecologia i Ciències Ambientals
Institut de Recerca de la Biodiversitat (IRBio)

Universitat de Barcelona

http://www.ub.es/softevol/domino

May 10, 2016

UNIVERSITAT DE BARCELONA

EG&B Evolutionary Genomics & Bioinformatics

1 Overview

The development of molecular markers is one of the most important challenges in phylogenetic and genome wide population genetics studies, especially in non-model organisms. A highly promising approach for obtaining suitable markers is the utilization of genomic partitioning strategies for the simultaneous discovery and genotyping of a large number of markers. Unfortunately, some of these markers may not provide enough information to solve specific evolutionary questions.

We have developed DOMINO, a bioinformatics tool for informative marker development from both NGS data and pre-computed sequence alignments. The application implements popular NGS tools with new utilities in a highly versatile pipeline specifically designed to discover or select personalized markers at different levels of taxonomic resolution.

Availability and implementation: DOMINO is an open source and multiplatform software that uses Perl as the main scripting language for the new implemented functions and the Qt framework for the graphical user interface. The software is freely available from www.ub.edu/softevol/domino.

Authors

Cristina Frías-López	cristinafriaslopez@ub.edu
Jose Francisco Sánchez-Herrero	jfsanchezherrero@ub.edu
Miquel A. Arnedo	marnedo@ub.edu
Alejandro Sánchez-Gracia	elsanchez@ub.edu
Julio Rozas	jrozas@ub.edu

DOMINO Publication

Frías-López, C.*, Sánchez-Herrero, J. F.*, Guirao-Rico, S., Mora, E., Arnedo, M. A., Sánchez-Gracia, A. and Rozas, J. 2016. DOMINO: Development and selection of informative molecular markers for studies in non-model organisms. *In Preparation*.
*, equal contribution

DOMINO Web Site

www.ub.edu/softevol/domino

2 Installation

DOMINO is distributed as compressed archives, which includes all files needed to install/run the software (source code, executable binaries and example data files, from the DOMINO website (www.ub.edu/softevol/domino).

DOMINO Distribution Package

The distributed package contains the following files:

Installer and Pre-compiled versions

DOMINO_version_OSX_Installer.dmg. Installer for Mac OS X operating systems.

DOMINO_version_OSX_Compiled.zip. Pre-compiled version for Mac OS X operating systems.

DOMINO_version_LinuxDistribution_Installer.run. Installer for Linux distributions.

DOMINO_version_LinuxDistribution_Compiled.rar. Pre-compiled version for Linux distributions.

Source code, example files and documentation

In the DOMINO Web page we have also included the source code, some example files and the documentation of the program.

DOMINO_Repository.zip. Compressed folder including all files needed to compile the software:

- /**docs**. Folder that includes documentation and some miscellaneous information.

- /**example**. Folder with the 4 FASTQ files of the example data set.

- /**src**. Folder that includes the source code of DOMINO. The DOMINO_Qt_code folder includes the Qt and the C++ code, while the DOMINO_perl_code folder includes the new Perl scripts specifically developed for the project.

- /**files**. Folder with the DOMINO installation files, including the compressed files of Perl modules, executable binaries, and third party programs (already compiled). Some installers for Linux and Mac OS X systems are also included.

- **install.sh**: Shell script for installing the command-line version of DOMINO.

- **README and NEWS**. Text files with relevant information about DOMINO.

- **Change.log.** Text file containing the different updates of the DOMINO project.

- **LICENSE.txt. FALTA ESTO**

DOMINO Software Requirements

Mac OS Systems (Mac OSX 10.x.x or higher)

Perl programming language. Perl should be installed by default on Mac OS X operating systems. If not, please follow instructions at http://learn.perl.org/installing/osx.html.

C++ compiler. This compiler is included in the Xcode integrated development environment, which can be installed via App Store.

`zlib`. This compression library should be installed by default on Mac OS X operating systems.

Linux systems (tested in Ubuntu 10.04 LTS)

`Perl` programming language. Perl should be installed by default on Linux distributions. If not, please follow instructions at http://learn.perl.org/installing/unix_linux.html.

C++ compiler. This compiler is included in the build essential package or can be installed thought the system package management tool (`g++` compiler).

Ubuntu: `sudo apt-get install build-essential`

`zlib`. This compression library can be installed thought the system package management tool. It is also available at http://www.zlib.net/

Ubuntu: `sudo apt-get install zlib1g-dev`

Windows Systems

In preparation

DOMINO Installation

DOMINO complete version

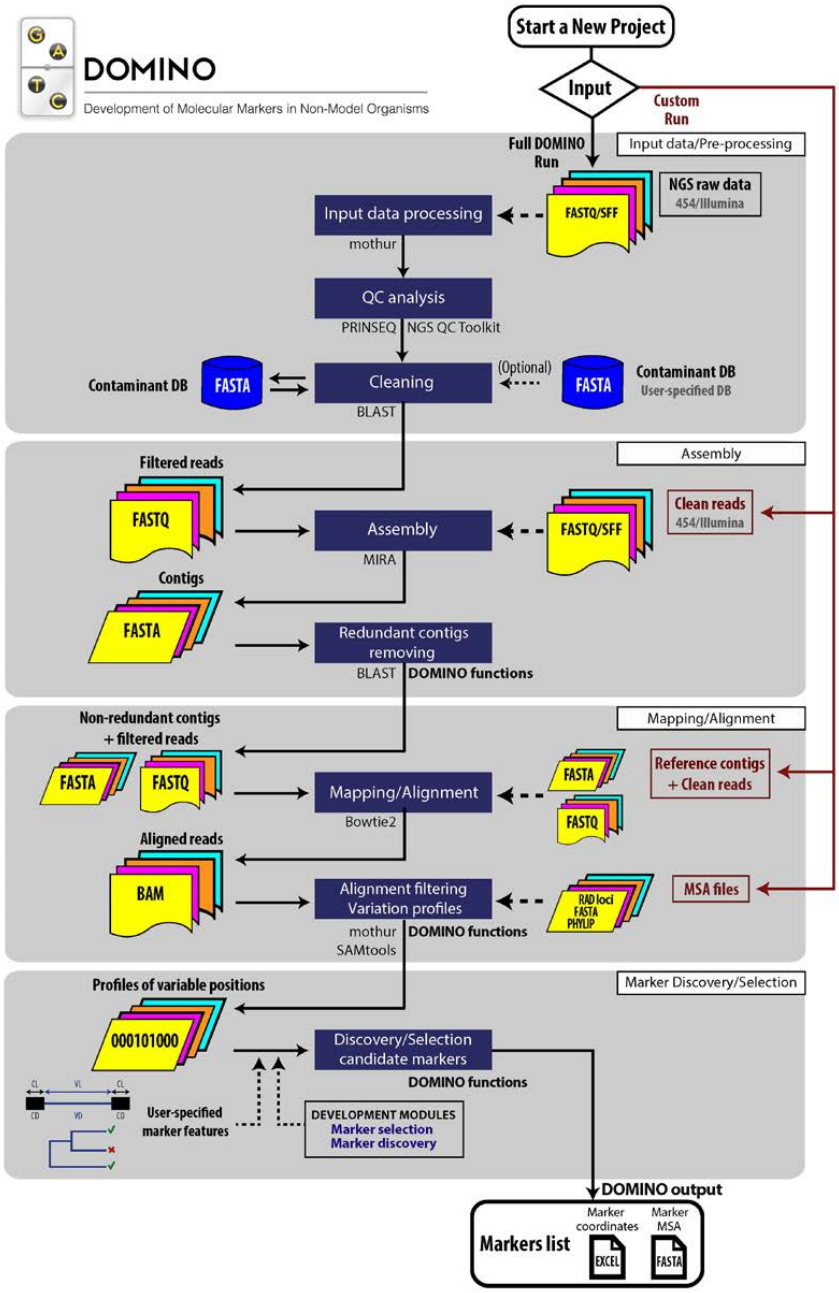
By executing the installers included in the DOMINO distribution package (`DOMINO_version_OSX_Installer.dmg/DOMINO_version_LinuxDistribution Installer.run`), all needed steps to install the software, including the creation of necessary folders and files as well as the Desktop icon, will be completed. After the installation, DOMINO can be run either under the GUI or under the command-line version.

DOMINO command-line version

In the command-line prompt (shell) run `sh install.sh` and follow the instructions. The installer will generate all folders and files necessary to run DOMINO under the command line version in Mac and Linux systems. All DM `Perl` scripts as well as the executable files of the external software integrated in the pipeline (third-party software) will be included within the directory `/bin`.

3 DOMINO –Workflow & Quick Guide

The DOMINO GUI consists in six TABs (GUI windows) that must be successively applied (under the Full DOMINO Run), or selectively skipped (under the different Custom Run options). The workflow behind schematizes the principal running steps across the four main phases.



1) Start a New Project

This TAB provides DOMINO information on the location of the relevant directories: the `Perl` and the DOMINO directories as well as other general options.

2) Input Data / Pre-Processing

Use this TAB to input your NGS data files. These files must contain long (454) or short (Illumina single or paired-end) raw data typically from a reduced partitioning protocol with several taxa (taxa panel). The sequences of each taxon should be properly barcoded (aka, tag, MID or index) or entered in separate files. In this phase**,** DOMINO will conduct all needed read pre-processing steps (i.e., cleaning and trimming of low quality, low complexity, short length and contaminant reads). In this window the user can configure some of the pre-processing parameters.

Scripts and software: `Mothur`, `PRINSEQ`, `NGS QC Toolkit`, `BLAST`, in-house `Perl` scripts.
Input: Raw reads
Output: Pre-processed reads

3) Assembly

This TAB allows setting some of the parameters for the *de novo* assembly phase, which will be performed separately for each taxon. After the assembly, DOMINO will remove repetitive contigs and reads for further steps.

Scripts and software: `MIRA`, `BLAST`, in-house `DOMINO Perl` scripts (`DM` scripts).
Input: Pre-processed reads
Output: Non-redundant contigs and unassembled reads

4) Mapping / Alignment

This TAB allows specifying the relevant parameters to obtain the profiles of variable sites between pairs of taxa (pairwise profiles). Here, the user can also enter their own pre-processed reads, reads plus reference sequences or multiple sequence alignments (MSA).

DOMINO will map the pre-processed reads to the contigs from the assembly phase or to the supplied reference sequence(s) and will perform the post-mapping quality-filtering and a conservative variant calling.

Scripts and software: `Bowtie2`, `Mothur`, `SAMtools`, and `DM` scripts.
Input: Pre-processed reads and reference sequence(s); MSA in various formats (see below for details).
Output: Profiles of variable sites across taxa

5) Marker Discovery / Selection

This TAB allows the user to select the taxa and the parameters values to be used for discovering or selecting the desired informative markers.

Scripts and software: `DM` scripts.
Input: Profiles of variable sites across taxa
Output: list of designed/selected informative markers

6) Files Viewer

This TAB shows the relevant files and directories used in each DOMINO run, and the list of the informative markers identified or selected with their coordinates (in contigs or reference sequence(s)).

EXAMPLE DATA FILE

To familiarise the user with the DOMINO GUI, we have included in the distributed package an example dataset (`4000Nemesia_fastq_files.zip`), of a panel with four taxa (which is a small subset of the RRL reported in Frías-López *et al.* 2016). This data set is used throughout this

manual to illustrate the different phases of the DOMINO workflow. It includes raw data (a subset of 4,000 reads from a 454 sequencing of an RLL experiment; 2.1 Mbp total; average read length of 518 bp; N50 of 611 bp) in four FASTQ files (one FASTQ file per taxon: `N_raripilia061.fastq`; `N_raripilia079.fastq`; `Nemesia_sp043.fastq`; `I_brauni098.fastq`).

Computational requirements issues

An important point to be considered before using DOMINO with NGS data (especially in the Pre-processing and Assembly phases) is the raw data size. Although the bioinformatics tools included in DOMINO can perfectly handle these kinds of data, they can consume substantial amounts of computational resources, especially RAM memory. We do not recommend applying the Full DOMINIO run with data sets of many millions of short reads in a typical desktop computer (with an insufficient number of CPUs and limited RAM); this process might take a very long (unacceptable) time to complete, or even cause a system crash.

For massive NGS data sets (typically more than 5 million short reads per file), the user can either run DOMINO under the command line version using high performance computers (i.e., a computer cluster with high large amounts of CPUs and lots of RAM and hard disk space) or, take advantage of the GUI Custom run options and enter DOMINO partially processed data, e.g. pre-processed reads, pre-assembled contigs or alignment files (SAM/BAM) obtained from other more memory-efficient software.

Using the command-line option

The user can also run DOMINO under the command-line prompt. This option allows running domino in high performance computer clusters and managing some extra options and parameters than the GUI version, such as a second (optional) iterative assembly round with CAP3 (using the contigs and singletons obtained from MIRA as input data), or activate extended options in the Alignment/Mapping and Marker Discovery/Selection phases. See the DOMINO website to find more information of the basic DOMINO command-line option, as well as for the default parameter values used in the GUI.

3 DOMINO ––START Your Project

Set the Perl and DOMINO Path boxes

In this first TAB, the user should specify the path to the folder with the `Perl` scripting language, either selecting the default `Perl` executable file (Default Perl path, usually recognized in all UNIX-based systems) or entering a different path name (Enter Perl executable path). The DOMINO path project, which will be used to write the output files, can also be specified here.

Miscellaneous Parameters box

In this box, the user is asked for the number of processors (cores) to be used for computation and whether the intermediate files will be kept or eliminated.

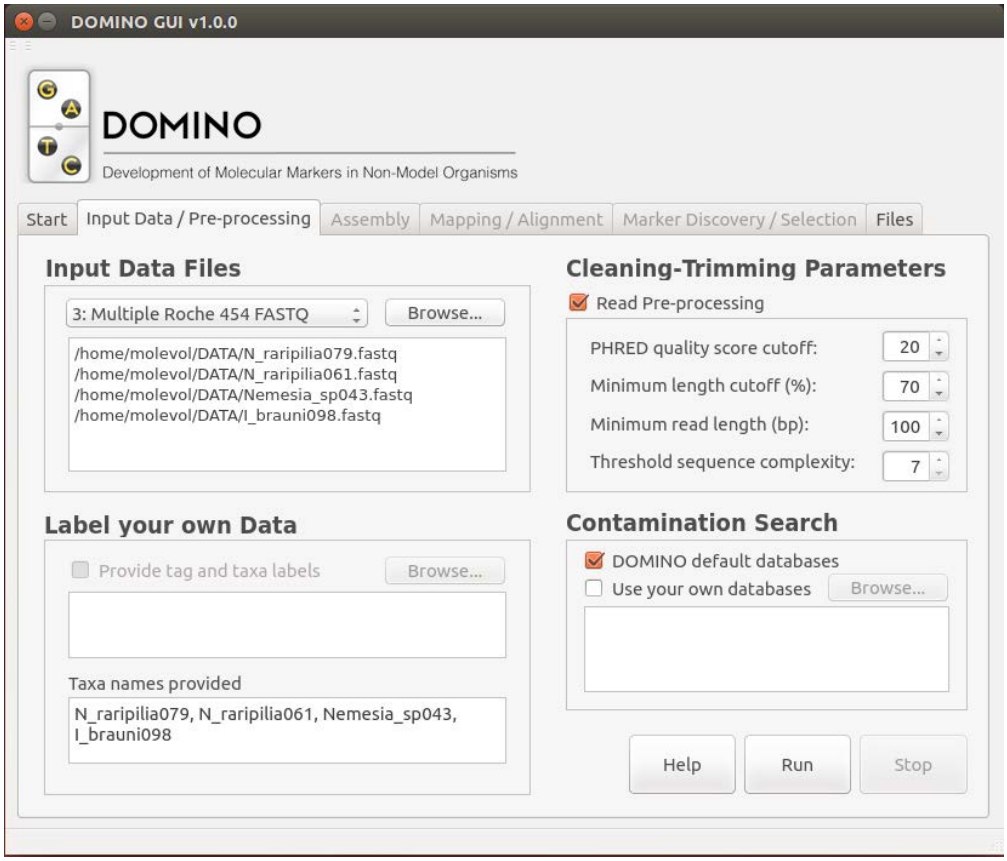
Full DOMINO Run button

This button starts a complete DOMINO run (that is, performing consecutively all core steps of the four DOMINO phases), from an input with raw NGS data to the final listing of candidate markers.

Custom Run button

Use this button to skip some DOMINO phases. Under this option, users can load their own cleaned reads, reference sequence(s) or pre-computed MSA. This option also permits resume a previous DOMINO execution.

4 DOMINO ––Input Data/Pre-Processing



Input Data box

DOMINO accepts different types of NGS input data files. The program can handle 454-SFF files and FASTQ files with 454, Illumina single or paired-ends raw reads (in various formats). The raw data of each taxon can be entered separately, or combined in a single file (such as in many SFF files). In this case, DOMINO requires that the DNA sequence data of each taxon was appropriately labelled (tagged with barcodes -MIDs).

Note that in a full run from short reads, DOMINO applies an assembly-based approach; the program is therefore optimized to work with genome partitioning methods in which the length of the size-selected or enriched fragments and the sequencing depth are enough to permit the assembly of putative homologous fragments. For data from other sequencing approaches (RAD-based data, such as. RAD-Seq, ddRAD or GBS) see the Mapping/Alignment and Marker Discovery/Selection sections of this manual.

Accepted data types

1: Roche 454 SFF

A single file in the Standard Flowgram Format (SFF), containing all 454 reads of all taxa from the panel, accordingly tagged (with MID tags). File Extension: `*.sff`

2: Roche 454 FASTQ
A single file in FASTQ format, containing the 454 reads of all taxa from the panel together, accordingly tagged (with MID tags). File Extension: `*.fastq`

3: Multiple Roche 454 FASTQ
Multiple FASTQ files, each file should contain the 454 reads of each taxon from the panel. File Extension: `*.fastq`

4: Illumina single-end FASTQ
A single file in FASTQ format, containing Illumina single-end reads of all taxa from the panel, accordingly tagged (with MID tags). File Extension: `*.fastq`

5: Multiple Illumina single-end FASTQ
Multiple single-end FASTQ files, each file should contain the raw illumina single-end reads of each taxon from the panel. File Extension: `*.fastq`

6: Two Illumina paired-end FASTQ
Two FASTQ files, one file should contain the left "_R1", and the other the right "_R2" fragment ends from a paired-end Illumina experiment, sequencing all taxa from the panel. Each taxon should be appropriately tagged (with MID tags). File Extension: `*.fastq`

7: Multiple Illumina paired-end FASTQ
Multiple paired-end FASTQ files. In this case the user should provide two illumina files of each taxon from the panel, one for the left "_R1" and another for the right "_R2" fragment ends. File Extension: `*.fastq`

Filenames
Structure:
`File_name[_Rn].extension`, where:
`File_name` stands for the taxon identifier (periods, commas or blank spaces are not allowed).
`[_Rn]` is optional and it is used to indicate the left "_R1" or right "_R2" fragment ends of a paired-end sequencing experiment.
`extension` stands for data type (SFF or FASTQ files).

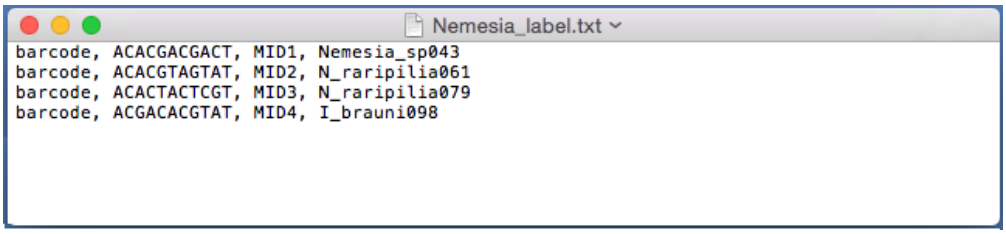
Examples:
`I_brauni098.fastq`, a FASTQ file with single-end reads from the *I_brauni098* taxon.
`Dmelanogaster_R1.fastq`, a FASTQ file with the left reads (paired-end data) of *Dmelanogaster* taxon.
`Dmelanogaster.fastq`, a FASTQ file with reads of *Dmelanogaster* taxon.

Taxon names
Taxon name lengths must be less than 25 characters. In case of larger filenames, you can rename your files including the tag `id-` to indicate the part of the filename that DOMINO will use as taxon name.

Example:
`FileXXX236363663id-Dmelanogaster_R1.fastq` --> taxon name: *Dmelanogaster*
`File1234-id-Dmelanogaster.fastq` --> taxon name: *Dmelanogaster*

Label your Data box

Since DOMINO accepts different types of input data files, it might require extra information in some cases. When the input file is a single 454 SFF or 454/Illumina FASTQ file, the program will request information on the particular nucleotide sequence used as MIDs or, alternatively, the names of the taxa in the panel (Provide a Tag & Taxa labels option). The user must provide this information in a text file (see the example below; file: `Nemesia_label.txt` the DOMINO website.



Loading separate input data files (e.g., multiple FASTQ files), each one with information from a single taxon, DOMINO will use the left part of the filename as the taxon name, excluding, if any, the "_R1" or "_R2" labels, which are used to indicate the left or right reads of a paired-end. For instance, taxon name extracted from the filename `I_brauni098.fastq`, will be *I_brauni098*, while `Dmelanogaster_R1.fastq`, `Dmelanogaster_R2.fastq` and `Dmelanogaster.fastq` will have the same specific taxon name (*Dmelanogaster*).

Cleaning-Trimming parameters box

Use this box to set the values for read pre-processing parameters. Nucleotides with quality values lower than the specified *PHRED quality score cut-off*, the reads with a % of valid nucleotides lower than the selected Minimum length cut-off, the reads shorter than the stablished Minimum read length, and the reads with Threshold sequence complexity values lower than those pre-defined by the user, will not be used in further steps.

Contamination Search box

This box allows selecting the database for the contaminants filtering step. DOMINO includes the UniVec database (which include information of vectors, adapters, linkers or other cloning contaminants), and the genome sequence of some prokaryotic (including *E. coli*) species and some virus as default databases for performing this task. The user can also load their preferred database (Use my own databases option) in FASTA format (see the file `MyOwnContaminantDB.txt` in the DOMINO website as an example).

By default DOMINO databases include the following prokaryotic species databases:
* *Escherichia coli* BL21(DE3) chromosome, complete genome
gi: 387825439; ref: NC_012971.2

* *Pseudomonas aeruginosa* M18 chromosome, complete genome
gi: 386056071; ref: NC_017548.1

* *Saccharomyces cerevisiae* S288c chromosome I, complete sequence
gi: 330443391; ref: NC_001133.9

* *Staphylococcus aureus* subsp. aureus 6850, complete genome
gi: 537441500; ref: NC_022222.1

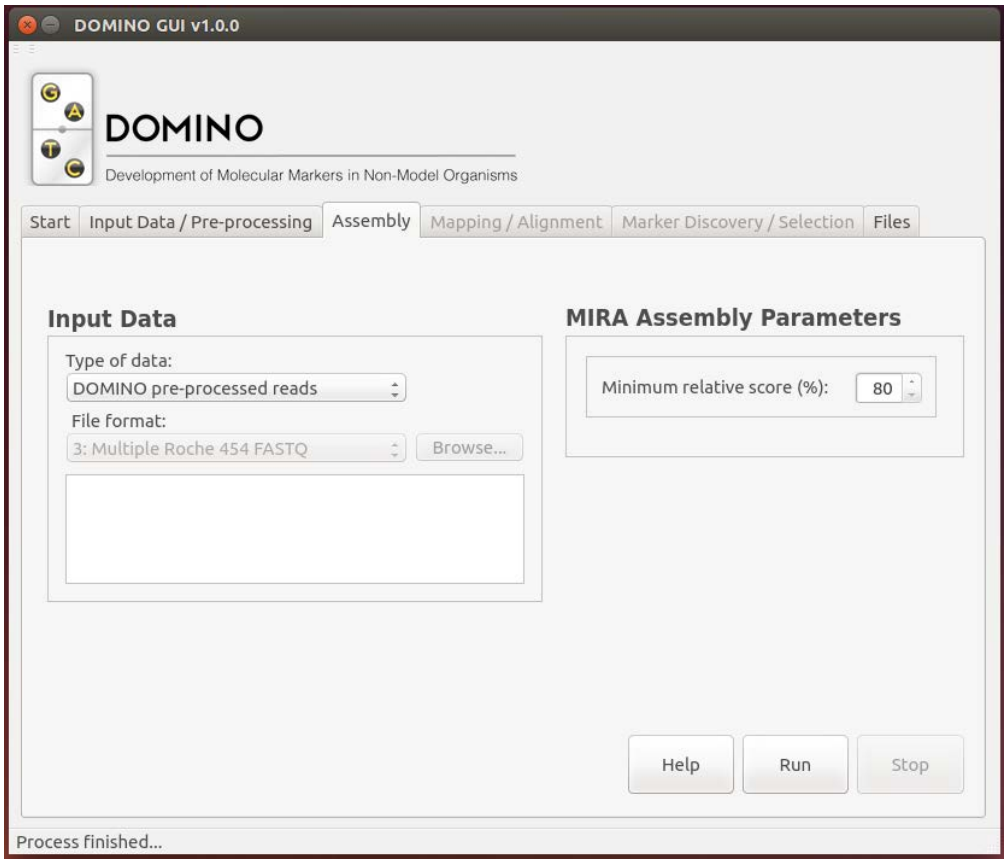
Results of our example data file

Using the default parameter values on the example data set (4,000 raw reads), DOMINO will select a final set of 3,981 high quality pre-processed reads.

Additional Information

The pre-processing steps are performed by using several scripts and software, such as `Mothur`, `PRINSEQ`, `NGS QC Toolkit`, `Blast` and several new functions implemented in an in-house written `Perl` script (`DM_Cleaning.pl` PERL script). Please read the documentation provided by these software and their description in the DOMINO paper (Frías-López *et al.* 2016) for details on parameters and options. See the DOMINO website to see some examples of the input file formats accepted by DOMINO.

5 DOMINO ––Assembly



DOMINO performs the *de novo* assembly of pre-processed reads separately for each taxon from the panel by using MIRA. DOMINO identifies the reads encompassing repetitive regions (reads classified as HAF6, HAF7 and MNRr in MIRA; see the MIRA documentation for details), and remove them from further DOMINO steps. Later, and to avoid including redundant contigs in the next steps, DOMINO conducts an all-vs-all contigs BLAST search. Contigs involved in positive hits with low e-values (cut-of E-value of 10^{-50}), with an overlapped region higher or equal than 85% (of the shorter contig length), and minimum similarity value of 85% (of the overlapped region) are collapsed, i. e., only the longest contig of each pair in a positive Blast hit will be used as a reference sequence for the next Mapping/Alignment step.

Input Data box

In this box, the reads pre-processed in previous DOMINO executions (in a complete/standard Full DOMINO Run option) or the reads directly supplied by the user (under the Custom Run option) can be loaded to perform the assembly.

Type of data: Data types

There are two options:

DOMINO pre-processed reads

This is the default option. Use this option if you want to perform the assembly using the reads previously pre-processed by DOMINO (either in the current or in a previous DOMINO execution).

User-supplied pre-processed reads
Using the Custom Run option, the user can enter directly in the assembly phase, by skipping all previous pre-processing steps. In this case, the user must supply its own data files in one of the accepted formats through the *File format* option. The filename and taxon name structure must be as specified in the Filenames and Taxon names sections in the description of previous TAB. In order to avoid problems with read naming and pair-end nomenclature we strongly recommend to make use of the Input Data/Pre-processing TAB to prepare user-supplied input files for the DOMIMO Assembly phase (i.e., by running a DOMINO pre-processing step in the previous TAB with the cleaning-trimming and contamination search options deselected).

File format: Accepted data types
3: Multiple Roche 454 FASTQ
Multiple FASTQ files, each file should contain the raw 454 reads of each taxon from the panel. File Extension: `*.fastq`

5: Multiple Illumina single-end FASTQ
Multiple single-end FASTQ files, each file should contain the raw illumina single-end reads of each taxon from the panel. File Extension: `*.fastq`

7: Multiple Illumina paired-end FASTQ
Multiple paired-end FASTQ files. In this case the user should provide two illumina files of each taxon from the panel, one for the left "_R1" and another for the right "_R2" fragment ends. File Extension: `*.fastq`

The filename structure is given in the previous TAB (Input Data/Pre-processing TAB).

MIRA Assembly Parameters
In this box, the user can set the values of some relevant parameters for the *de novo* assembly phase with MIRA, such as the minimum % matching for the assembly of two reads (Minimum relative score). In case of a second (optional) CAP3 assembly, the user can specify the minimum overlapping length and the minimum % of identity accepted in these overlapping regions.
*Note***.** The default values for these parameters has been set accordingly with the type of data (e.g. they are different for 454 and Illumina data).

Results of our example data file
The MIRA assembly, using the default DOMINO values, of each of the four taxon in the example, will generate 100, 128, 134 and 145 contigs for *N_raripilia061*, *N_raripilia079*, *Nemesia_sp043* and *I_brauni098*, respectively (see also the Frías-López *et al.* 2016; supplementary Tables S4-S5).

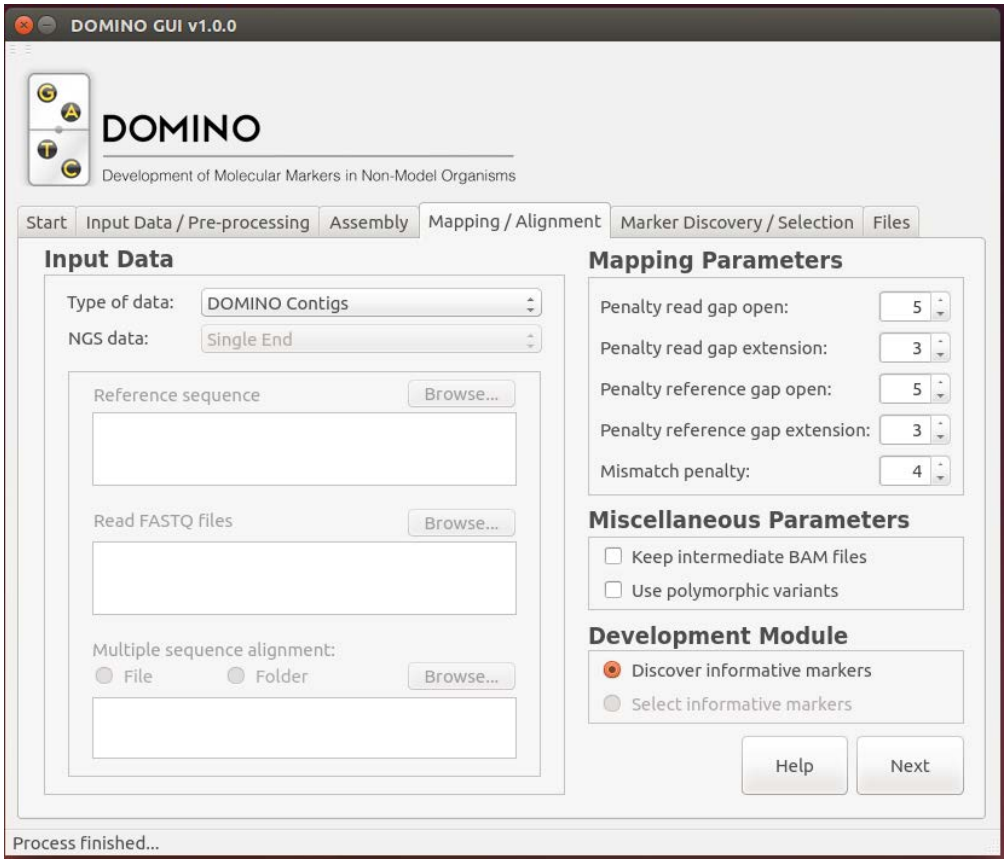
MIRA limitations

MIRA is based on a highly accurate overlap graph algorithm and usually shows good performance with short reads and contigs, as those expected to be handled by DOMINO (data from typical genome partitioning, low coverage or small genome size experiments). For the assembly of much larger sequences (which needs many millions of short reads) the overlap graph strategy becomes computationally intensive and should be either avoided or carried out in high performance computers (see the Computational requirements issues section7).

Additional Information

In addition to MIRA (and CAP3) software, DOMINO, uses a series of new developed functions implemented in an in-house `Perl` script (`DM_Assembly.pl` PERL script) for the assembly phase. Please read the documentation provided by these external software and their description in the DOMINO paper (Frías-López *et al.* 2016) for details on parameters and options. See the DOMINO website to see some examples of the input file formats accepted by DOMINO.

6 DOMINO ––Mapping/Alignment



At the end of Mapping/Alignment phase, DOMINO builds the arrays of variable positions between pairs of taxa required for further marker Discovery. This phase is performed in four different steps.

Mapping/Alignment
First, the pre-processed reads from each taxon are independently mapped back to the contigs (and singletons) obtained in the assembly phase using `Bowtie2`. The resulting BAM files [*n* x (*n*-1) files, *n* = number of taxa in the panel; four in our example] will contain all pairwise combinations of reads from one taxon aligned back to the contigs of all other taxa separately (i.e., ignoring the alignments between reads and contigs from the same taxon).

For instance, using the assembly reference sequence data of taxon#1, DOMINO will create 3 different BAM files (contigs of taxon#1 with reads of taxon#2, with reads of taxon#3, and with reads of taxon#4), and so on. Using the example dataset (a panel of four taxa), DOMINO will perform four separated assemblies, one per taxon; in the mapping step, the pre-processed reads of each of these four taxa will be aligned separately upon contigs of the other three taxa, that is, a total of 4 x 3 = 12 BAM/SAM files.

Filtering mapping errors & non-useful alignments
DOMINO applies some additional filtering steps to remove alignments with mapping errors, unmapped regions or multimapping reads. These problematic alignments can definitely generate

false informative markers (regions with artefactual levels of nucleotide diversity). Alignments with an unusually large number of mapped reads, which might result from repetitive regions, are also removed since they are not useful as sources of informative markers. For that, DOMINO estimates the average read coverage (*c*) among all alignments, and removes those with coverage equal or greater than a critical value ($P < 10^{-5}$; this value could be modified in the command-line version), which is obtained from a Poisson distribution with mean *c*. After these filters DOMINO builds a pileup file for each of the filtered BAM files (12 in the example data set), using the SAMtools suite.

Filtering sequencing errors & ambiguity codes
Since sequencing errors can severely affect marker identification and selection, DOMINO implements a very conservative variant calling function (see Frías-López *et al.* 2016 supplemental methods for details) for detecting and masking putative errors. First, to avoid the calling of spurious nucleotide variants in low sequencing coverage experiments (i.e. erroneously assigned variants fixed between the taxa panel), DOMINO masks the information from positions with only one read mapped to the reference.

DOMINO incorporates a similar conservative criterion to use only highly credible polymorphisms when the Polymorphic variants option is activated (see the Marker Discovery/Selection section). For positions with 8 or more reads mapped, DOMINO discards those polymorphic variants in which the frequency of the minor allele is significantly lower than the expected for a diploid individual ($P < 0.05$ in Binomial distribution with $p = 0.5$), likely corresponding to a sequencing error. For lower coverage values, DOMINO will use the information of a polymorphic variant only if the frequency of the minor allele is present in two or more mapped reads.

DOMINO also checks the presence of positions with ambiguity codes (generated by MIRA) and decides whether they are variable depending on the nucleotides present in the rest of positions. For instance, if in the reference sequence appears a "Y" (IUPAC ambiguity code for "C/T"), and all nucleotides in the mapped reads have an "A" for this position, DOMINO considers this position as variable. On the other side, if all reads have a "C" in this position, DOMINO considers this position as invariable.

Profile of variable sites between pairs of taxa
At the end of the mapping step, DOMINO builds a single profile of variable sites (merged profile), combining the information from all pairwise pileup having the same reference sequence. In the example data set, DOMINO will build four merged profiles (one for each assembled taxon). In case that user entered a single reference sequence, DOMINO generates only one merged profile.

Input Data box

The mapping step can be performed using the DOMINO files from the current project, or by loading directly the data files necessary to conduct the mapping step (under the Custom Run option). Users can also input their own pre-processed reads, appropriately pre-processed accordingly with the genome partitioning methodology used to generate the library and the NGS technology used for sequencing, and MSA files. In this case, and in order to avoid problems with read naming and pair-end nomenclature, we strongly recommend to make use of the Input Data/Pre-processing TAB to prepare user-supplied input files for the DOMIMO Mapping/Alignment phase (i.e., by running a DOMINO pre-processing step in the previous TAB with the cleaning-trimming and contamination search options deselected).

Type of data: Data types
DOMINO accepts different types of data files:

DOMINO contigs (Full DOMINO Run option)
Default value. Use this option if you want to use the contigs previously assembled in DOMINO, either in the current or in a previous DOMINO execution.

Multiple taxa references (Custom Run option)
Use this option to map the loaded pre-processed reads to reference sequences from several taxa (a separate reference sequence per taxon). The user must load two different kind of data files:

1) Reference sequences (such as size-selected or enriched library fragments or genome contigs and scaffolds) from multiple taxa. The user must upload *n* data files (being *n* the number of taxa in the panel) in multi-FASTA format, with the DNA sequence to be used as a reference in the mapping/alignment phase.
2) NGS data files. FASTQ files (one file per taxon for single-end reads; two files for paired-end reads) with the sequence of the pre-processed reads to be aligned to the supplied reference sequences. DOMINO will therefore conduct *n*(*n*-1) mapping/alignment steps. In addition, the user must indicate whether the supplied reads are from a single-end or paired-end sequencing experiment.

Single taxon reference(s) (Custom Run option)
Use this option to map the loaded pre-processed reads from one or more taxa to a single taxon reference sequence. This option is identical to the References from multiple taxa option, but using only a single reference sequence for all taxa. The user must load two kind of data files:

1) A reference sequence (such as size-selected or enriched library fragments or genome contigs and scaffolds) from a single taxon. The user must upload one data file (in multi-FASTA format) with the DNA sequence to be used as a reference in the mapping/alignment phase.
2) NGS data files. FASTQ files (one file per taxon for single-end reads; two files for paired-end reads) with the sequence of the pre-processed reads to be aligned to the supplied reference sequence. DOMINO will conduct *n* mapping steps, being *n* the number of taxa in the panel. In addition the user must indicate whether the supplied reads are from a single-end or paired-end sequencing experiment.

MSA file(s) (Custom Run option)
Use this option to identify informative molecular markers directly in one or more MSA of nucleotide sequences; each MSA may include DNA sequence information from size-selected library fragments (from a genome partitioning scheme) or from genomic contigs or scaffolds. Using this option, DOMINO will skip the mapping phase and use directly these MSA for the next step (Marker Discovery/Selection TAB). The user can load the MSA in any of the following formats.

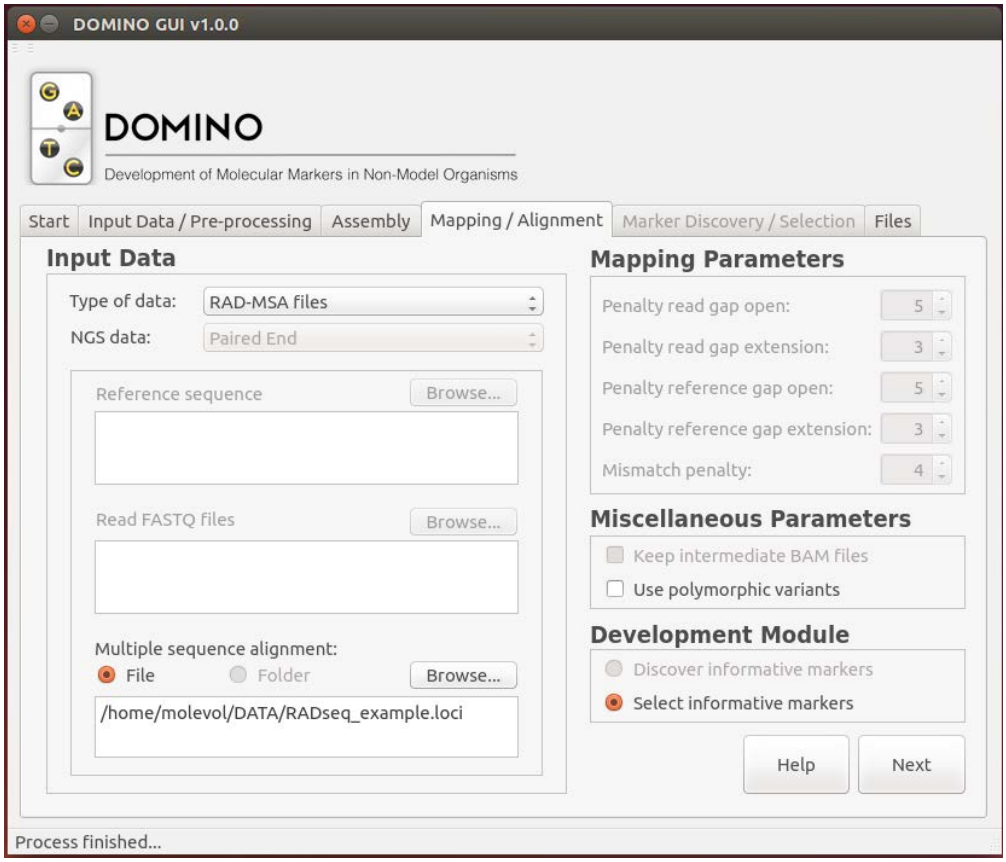
1) Various MSA files in PHYLIP or FASTA format. Each file must include only one MSA corresponding to a single library fragment or genomic region.
2) A multi-MSA file in PHYLIP format. A single data file with multiple MSA, each one in PHYLIP format. Each region are separated by the standard first-line PHYLIP identifiers (two numbers: the number of taxa, and the number of nucleotides).

RAD-MSA files (Custom Run option)
Use this option to select the most informative MSAs (RAD loci) among a supplied set, which must to be previously generated by some RAD software tools (e.g. PyRAD or Stacks software for RAD-Seq analyses). Using this option, DOMINO will skip the mapping phase and use directly these MSA for the next step (Marker Discovery/Selection TAB). The user can load the MSA in any of the following formats:

1) A multi-MSA file in PyRAD output loci format (the output file from PyRAD with the extension `*.loci`; see the PyRAD documentation for details). A single data file with all

MSA (one per RAD loci) from a RAD-Seq or a similar methodology. Each MSA must to be in FASTA format and separated by the character: `//`.

2) A multi-MSA file in Stacks output FASTA format (the output file from Stacks with the name `batch_X.fa`; see the Stacks documentation for details). A single data file with all MSA (one per RAD loci) from a RAD-Seq or a similar methodology. The sequence of the two haplotypes of each individual is included in each MSA.



Filenames and taxon names

The data files for the same taxa should have the same taxon id-name. The accepted filename structures are:

`[xx]id-yyy.fastq` --> Clean reads (FASTQ format) for taxon *yyy*

`[xx]id-yyy[_Rn].fastq` --> Clean paired-end FASTQ reads for taxon *yyy*. `_Rn`, indicate the left "_R1" or right "_R2" reads of a paired-end sequencing experiment

`[xx]id-yyy.contigs.fasta` --> Contigs for taxon *yyy* (FASTA format)

`[xx]id-yyy.fasta` --> Reference sequence (scaffolds; complete genome) for a particular taxon of the panel (FASTA format)

Where:

`xx`, could be any character or none

`yyy`, is the desired taxa name.

Therefore, some correct filenames could be:

`id-HomoSapiens.fastq`

```
123id-Dmelanogaster_R1.fastq
1id-Nemesia.contigs.fasta
Myid-Buthus.fasta
```

The accepted filenames structure and extension for the MSA of RAD loci are:
`Filename.loci` --> Multi-MSA in PyRAD loci output format
`Filename.fa` --> Multi-MSA in Stacks FASTA output format

Mapping Parameters (`Bowtie2` software) box
In this box, the user can modify penalty parameters for the `Bowtie2` mapping. Please read `Bowtie2` documentation.

Miscellaneous Parameters
Skip DOMINO mapping step. This box allows skipping the mapping step, and use mapping data from a previous DOMINO session.
Skip DOMINO MSA-parsing step. This box allows skipping the generation of variation profiles, and use profiles data from a previous DOMINO session.

Development Module
Use this box to select the specific DOMINO module to be used in the Discovery/Selection phase.

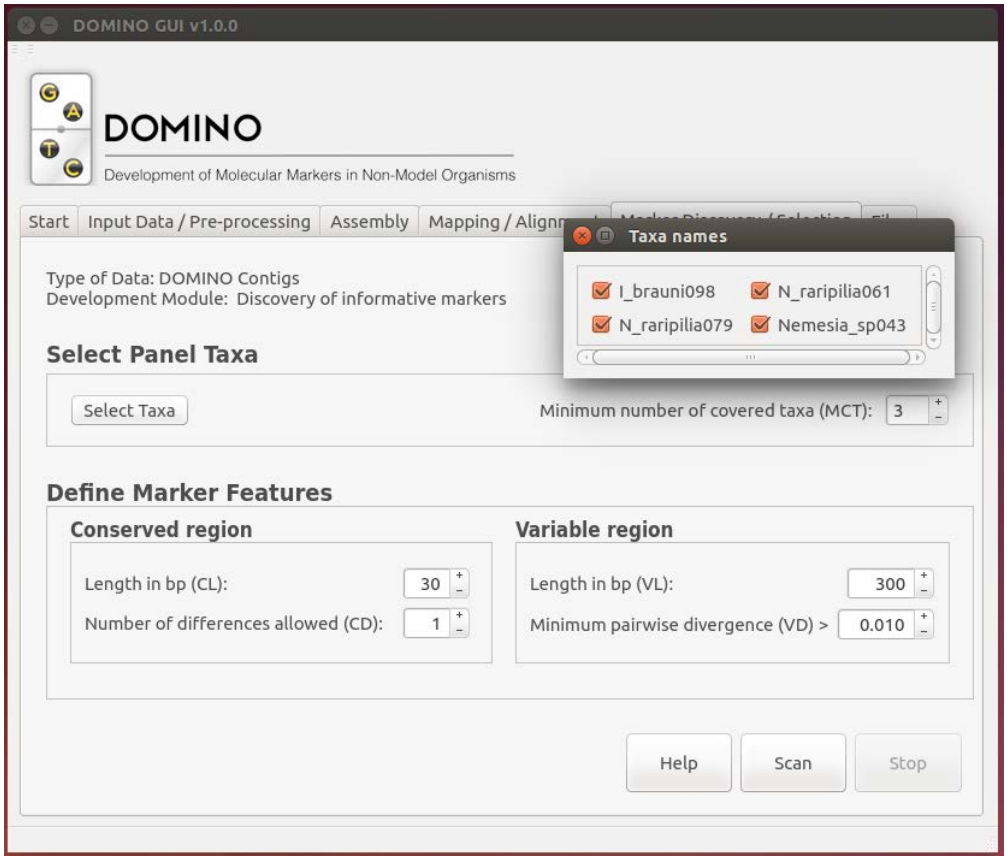
DOMINO marker discovery module
Under this module, the program will search for the presence of candidate marker regions (using a sliding window approach) across either the merged arrays of variable sites generated in current Mapping/Alignment phase or a set of pre-computed MSA loaded by the user using the MSA file(s) option in the Input Data box (Type of data).

DOMINO marker selection module
If the user chooses this module, the program uses an internal function to select the markers with the desired features among a set of pre-computed MSA. These MSA can be loaded using the MSA file(s) or the RAD-MSA files options in the Input Data box (Type of data).

Additional Information

The mapping and the building of the profiles of variable positions are performed using `Bowtie2` software, `SAMtools` suite and new developed functions implemented in a Perl script (`DM_MarkerScan.pl` PERL script). Please read the documentation provided by these software and their description in the DOMINO paper (Frías-López *et al.* 2016) for details on parameters and options. See the DOMINO website to see some examples of the input file formats accepted by DOMINO.

7 DOMINO ––Marker Discovery/Selection



DOMINO searches for informative markers (i.e., regions of a particular length with a desirable level of nucleotide variation among selected taxa or encompassing a minimum number of them (which can be optionally flanked by two highly conserved regions), in the merged arrays of variable sites generated in the previous phase.

Since DOMINO can search markers based on profiles built using different reference sequences (e.g. four in our example, one for each taxon), the same region can be identified/selected as a marker more than once. To avoid reporting this redundant information, DOMINO uses BLAST to identify and collapse these redundant markers across the different profiles of variable sites and report only one of them.

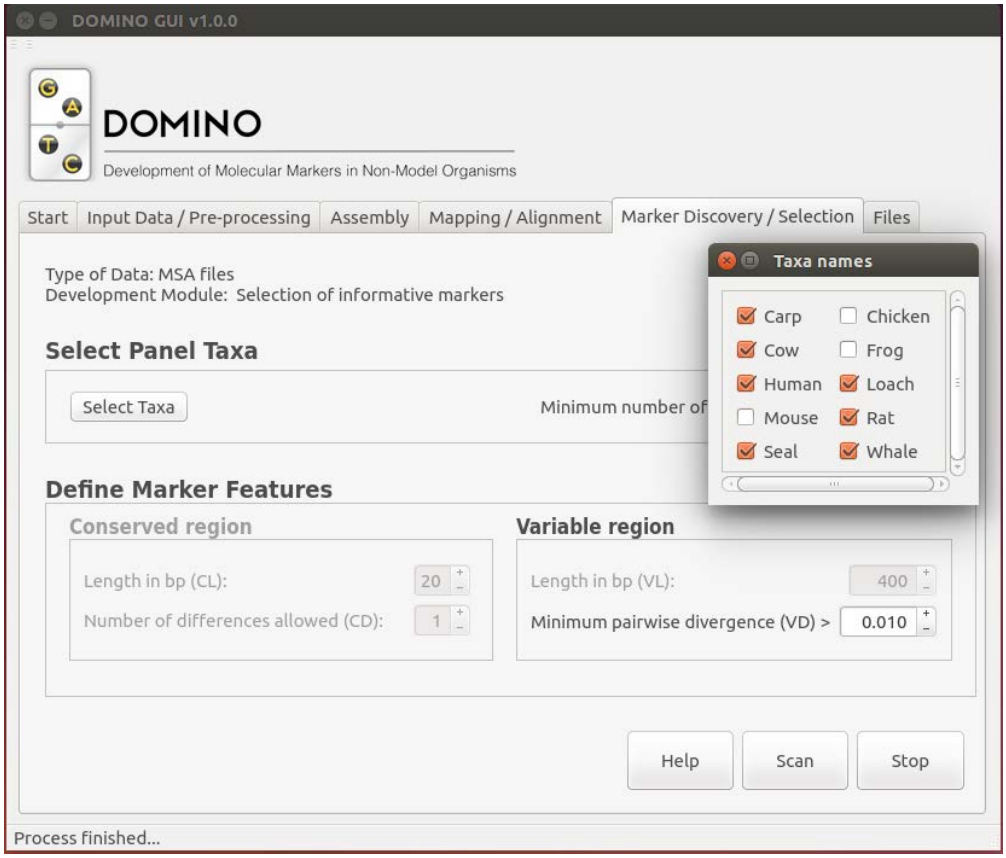
Panel Taxa box

Regardless of the number of taxa included in the input data file (taxa panel), the user can select in this box the taxa to be used in the Marker Discovery/Selection phase, and therefore to restrict the marker search to all or a subset of them (with a minimum of two). Moreover, DOMINO allows specifying the minimum number of taxa (amongst the currently selected taxa) that must satisfy the specified conditions (marker features) for a region to be considered as an informative marker (Minimum number of covered taxa, MCT). For instance, a value of MCT = 4 means that the user will restrict the marker search to aligned regions covered with information from all four taxa (as in the example). When the objective is to design markers useful for further PCR amplification and

sequencing in a larger focal taxa set (taking into account the phylogenetic relationships among the four taxa from the panel), and DOMINO detects few markers, the user might relax the search conditions by changing the MCT value. For instance, setting the MCT value to 3, DOMINO will search for markers in regions covered by at least 3 of the 4 selected taxa, and in which at least 3 of them exhibit the minimum variation desired for the marker (see the description of VD parameter for details). In case of using RAD data, DOMINO will ensure that, in each RAD loci, the number of variable positions between at least 3 taxa is within the required range (see the description of VP parameter for details). Hence, this option is very useful to find markers informative to resolve the phylogenetic relationships between or among specific groups (i.e. markers informative at different phylogenetic/population genetics ranges).

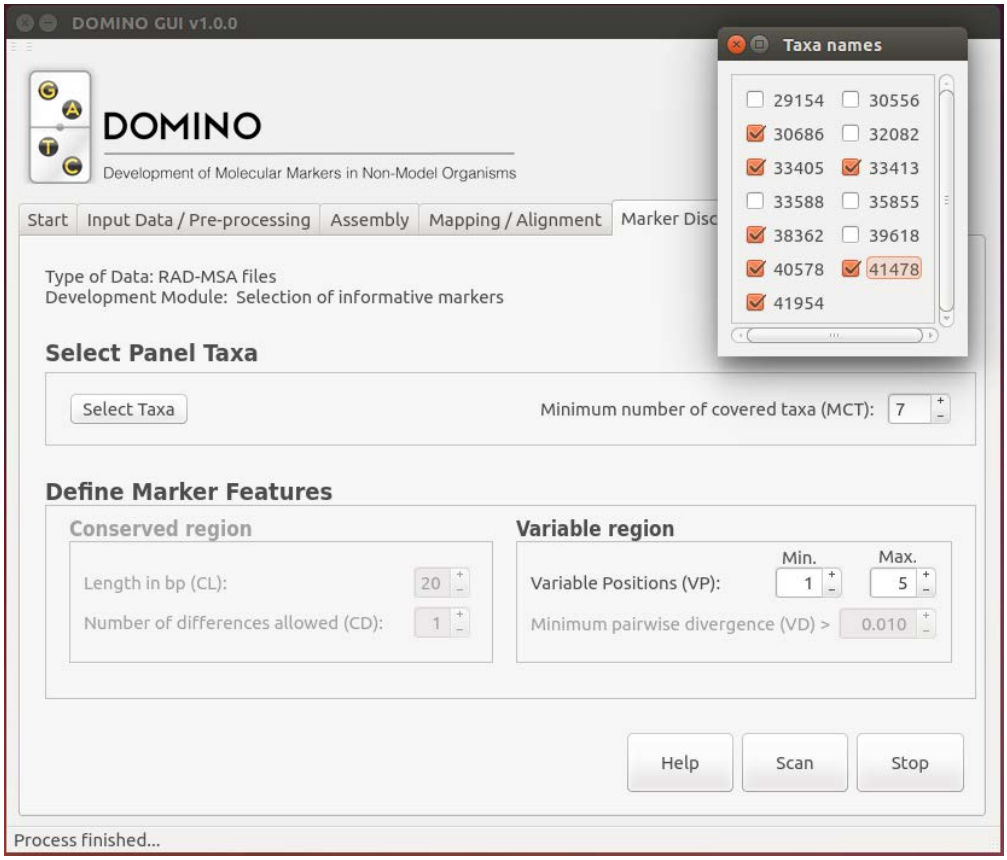
Define Marker Features

This box allows choosing a large number marker features. If the user are interested in obtaining informative markers for their further PCR amplification and sequencing in other phylogenetically related taxa (focal taxa), DOMINO will search for conserved stretches flanking the identified or selected marker; in this case the search can be restricted to conserved regions of a specified Length (CL), and exhibiting a maximum Number of nucleotide differences across taxa (CD). For the marker region itself, the user can search for a particular Length range (VL), and restrict the analyses to regions exhibiting a minimum level of variation (nucleotide substitutions per site) between any pair of taxa (VD value). Furthermore, using the Polymorphic variants option, DOMINO will use polymorphic positions to build the profile of variable sites.



RAD-MSA data and similar data

If the user enters RAD loci data (Custom Run option; see the Mapping/Alignment section), DOMINO allows selecting the most informative markers (RAD loci) in the same way and with the same definable features described above. In this case, however, a specific range of variable positions (VP) between the closest taxa instead of a minimum pairwise divergence level should be specified. The latter option will allow selecting informative RAD loci while excluding all cases exhibiting anomalous high levels of variation (which might reflect RAD tag clustering errors).



Scan / Re-Scan button

Use the scan button to start the Marker Discovery/Alignment phase. The Re-Scan button will be automatically activated after finished the first run; this button allows the user change some maker feature parameters to repeat the search analysis.

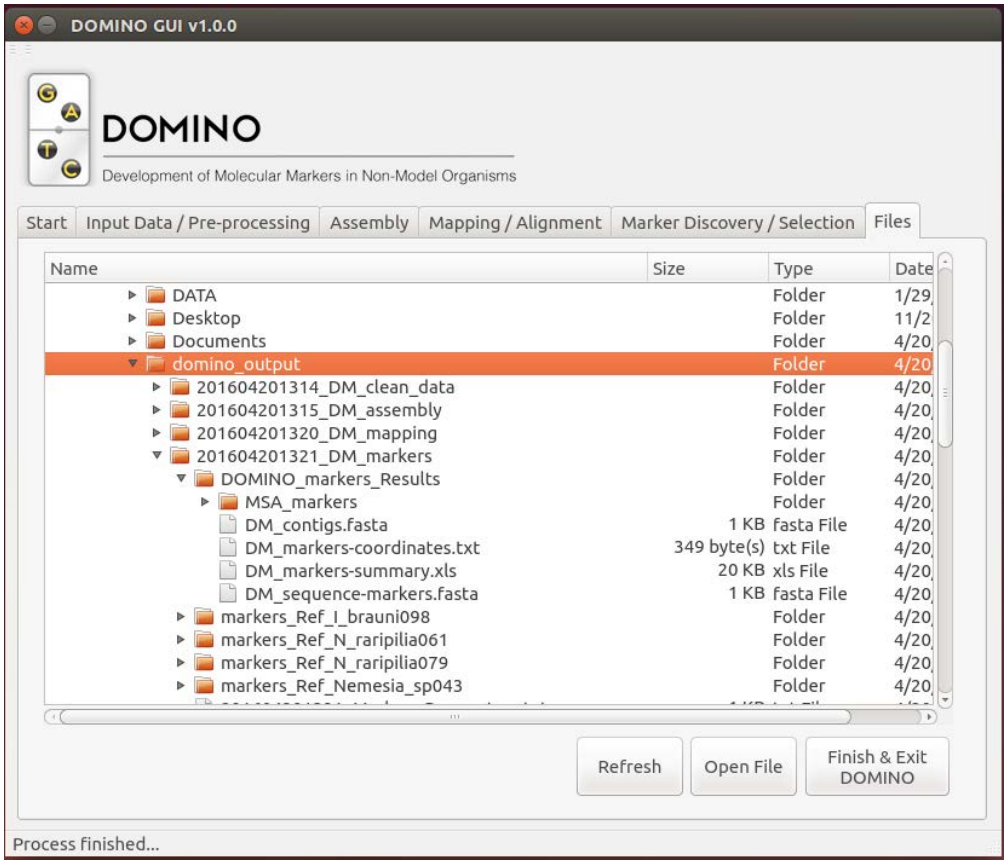
Results of our example data file

Under the Full DOMINO Run option and using the default parameter values for the previous tabs, DOMINO identifies 16 markers in 10 independent contigs (CL=30, CD=1, VL=300, VD=0.01, MCT=3); see the `4000Nemesia_Example_output.xls`, in the DOMINO website. See also the Frías-López *et al.* 2016 (supplementary Tables S4-S5) for the outcome for other parameter combinations.

Additional Information

The marker discovery and selection phase is mainly performed by our new developed functions implemented in the Perl script `DM_MarkerScan.pl`, although it also uses other scripts and pieces of software such as BLAST. Please read the paper describing DOMINO (Frías-López *et al.* 2016), for details on parameters and options. See the DOMINO website to see some examples of the input file formats accepted by DOMINO.

8 DOMINO --OUTPUT



In this TAB the user can easily visualize and get access to the folders and files that contains the results and all the stuff form intermediate analyses.

List of informative markers (`DOMINO_markers_Results` folder)

`DM_markers-summary.txt` and `DM_markers-summary.xls` contain (in plain text and Excel format) the relevant information about the markers discovered or selected by DOMINO. This information also includes, if requested, the coordinates of the conserved regions to be used for further PCR amplification and sequencing experiments.

The `DM_sequence-markers.fasta` file contains the DNA sequence of the complete region identified or selected as an informative marker in a multi-FASTA format (including the conserved regions if applicable). The `DM_contigs.fasta` file contains the DNA sequence of all contigs with one or more identified markers in a multi-FASTA.

In the sub-folder `MSA_markers`, DOMINO stores the MSA files of each identified or selected informative marker in FASTA format. In this case, the MSA will always contain only the marker region itself, regardless of whether the flanking conserved regions were requested or not. Furthermore, DOMINO also provides a file with all these MSA concatenated in FASTA format (markers are separated by a white space), which can be directly used for downstream phylogenetic or population genetic analyses.

9 Acknowledgements & Funding

This work was supported from the Ministerio de Economia y Competitividad of Spain (grants BFU2010-15484 and CGL2013-45211 to J.R., and CGL2012-36863 to M.A.A.), from the Comissió Interdepartamental de Recerca i Innovació Tecnològica of Spain (2009SGR-1287; 2014SGR1055; 2014SGR1604). J.R. and M.A.A. were partially supported by ICREA Academia (Generalitat de Catalunya), A.S-G. by a Beatriu de Pinós postdoctoral fellowship (Generalitat de Catalunya), C.F-L by an IRBio predoctoral fellowship (Universitat de Barcelona) and J.F.S-H by a FPU predoctoral fellowship (Ministerio de Educación y Ciencia).

4.2. The draft genome sequence of the spider *Dysdera silvatica* (Araneae, Dysderidae): A valuable resource for functional and evolutionary genomic studies in chelicerates

Presentamos el ensamblaje genómico de *Dysdera silvatica*, una araña nocturna perteneciente a un género que sufrió una importante radiación adaptativa en las Islas Canarias. El ensamblaje fue generado a partir de secuencias cortas de secuenciación masiva (*illumina*) y largas (*PacBio* y *Nanopore*). Nuestro ensamblaje de novo (1,36 Gb), representa un 80 % de lo estimado por citometría de flujo (1,7 Gb), está constituido por una elevada fracción (53,8 %) de elementos repetitivos interespaciados. La integridad del genoma, medida mediante el *software* BUSCO (*"Benchmarking Universal Single Copy Orthologs"*) y CEG (*"Core Eukaryotic genes"*), varía entre un 90-96 %, respectivamente. La anotación funcional basada tanto en predicción *ab initio* como en evidencias (incluyendo RNAseq de la propia especie) generó un total de 48.619 secuencias codificantes de proteínas, de las cuales, 36.398 (74,9 %) presentan dominios moleculares de proteínas conocidos o similitud de secuencia con la base de datos *Swissprot*.

El ensamblaje de *D. silvatica* supone el primer representante de la superfamilia Dysderoidea, y solo el segundo genoma disponible de Synspermiata, uno de los grandes linajes de arañas verdaderas (Araneomorphae). Las Dysderas, conocidas por sus numerosos casos descritos de adaptación a ambientes subterráneos, incluyen algunos ejemplos de especialización trófica en el grupo de las arañas y son excelentes modelos para el estudio de mecanismos de selección femenina críptica. Este recurso es por tanto muy útil como punto de partida del estudio de importantes cuestiones evolutivas y funcionales, incluyendo las bases de la adaptación a ambientes extremos y cambios ecológicos así como al origen y evolución de relevantes rasgos de las arañas como el veneno o la seda.

4

GigaScience, 8, 2019, 1–9

doi: 10.1093/gigascience/giz099
Data Note

DATA NOTE

The draft genome sequence of the spider *Dysdera silvatica* (Araneae, Dysderidae): A valuable resource for functional and evolutionary genomic studies in chelicerates

Jose Francisco Sánchez-Herrero [1], Cristina Frías-López [1], Paula Escuer [1], Silvia Hinojosa-Alvarez [1,2], Miquel A. Arnedo [3], Alejandro Sánchez-Gracia [1,*] and Julio Rozas [1,*]

[1]Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona (UB) and Institut de Recerca de la Biodiversitat (IRBio), Diagonal 643, 08028 Barcelona, Spain ; [2] Jardín Botánico, Instituto de Biología, Universidad Nacional Autónoma de México, Tercer Circuito Exterior S/N, Ciudad Universitaria Coyoacán, 04510 México DF, México and [3]Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals, Universitat de Barcelona (UB) and Institut de Recerca de la Biodiversitat (IRBio), Diagonal 643, 08028 Barcelona, Spain

[*]**Correspondence address**. Julio Rozas, Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona (UB) and Institut de Recerca de la Biodiversitat (IRBio), Diagonal 643, 08028 Barcelona, Spain. E-mail: jrozas@ub.edu http://orcid.org/0000-0003-4543-4577; Alejandro Sánchez-Gracia, Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona (UB) and Institut de Recerca de la Biodiversitat (IRBio), Diagonal 643, 08028 Barcelona, Spain; . E-mail: elsanchez@ub.edu http://orcid.org/0000-0002-6839-9148

Abstract

Background: We present the draft genome sequence of *Dysdera silvatica*, a nocturnal ground-dwelling spider from a genus that has undergone a remarkable adaptive radiation in the Canary Islands. **Results:** The draft assembly was obtained using short (Illumina) and long (PaciBio and Nanopore) sequencing reads. Our *de novo* assembly (1.36 Gb), which represents 80% of the genome size estimated by flow cytometry (1.7 Gb), is constituted by a high fraction of interspersed repetitive elements (53.8%). The assembly completeness, using BUSCO and core eukaryotic genes, ranges from 90% to 96%. Functional annotations based on both *ab initio* and evidence-based information (including *D. silvatica* RNA sequencing) yielded a total of 48,619 protein-coding sequences, of which 36,398 (74.9%) have the molecular hallmark of known protein domains, or sequence similarity with Swiss-Prot sequences. The *D. silvatica* assembly is the first representative of the superfamily Dysderoidea, and just the second available genome of Synspermiata, one of the major evolutionary lineages of the "true spiders" (Araneomorphae). **Conclusions:** Dysderoids, which are known for their numerous instances of adaptation to underground environments, include some of the few examples of trophic specialization within spiders and are excellent models for the study of cryptic female choice. This resource will be therefore useful as a starting point to study fundamental evolutionary and functional questions, including the molecular bases of the adaptation to extreme environments and ecological shifts, as well of the origin and evolution of relevant spider traits, such as the venom and silk.

Received: 6 May 2019; **Revised:** 27 June 2019; **Accepted:** 30 July 2019

Keywords: Araneomorphae; hybrid genome assembly; genome annotation; Canary Islands

Figure 1 Male of *Dysdera silvatica* from Teselinde (La Gomera, Canary Islands). Photo credit: Miquel Arnedo.

Data Description

Spiders are a highly diverse and abundant group of predatory arthropods, found in virtually all terrestrial ecosystems. Approximately 45,000 spider species have been recorded to date [1]. The nocturnal ground family Dysderidae ranks 17th out of 118 currently accepted spider families in number of species. The type genus of the family, *Dysdera* Latreille, 1804, includes half of the family diversity (282 species). This genus is remarkable in several aspects. First, it represents one of the few cases of stenophagy, i.e., prey specialization, across spiders [2]. Many species in the genus have evolved special morphological, behavioral, and physiological adaptations to feed on woodlice, including modifications of mouthparts, unique hunting strategies, and effective restriction to assimilation of metals into its tissues [3–7]. Because of their chemical defenses and ability to accumulate heavy metals from the soil, woodlice are usually avoided as prey by most spiders, including generalist *Dysdera* [2,4, 5,7]. Although mostly circumscribed to the Mediterranean region, *Dysdera* has colonized all the Macaronesian archipelagoes and has undergone a remarkable species diversification in the Canary Islands [8]. As many as 55 species have been recorded across the 7 main islands and islets of this archipelago, being most of them single-island endemics [9]. Although multiple colonization events may account for the initial origin of species diversity the bulk of this diversity is the result of *in situ* diversification [8]. *Dysdera* spiders have adapted to a broad range of terrestrial habitats within the Canary Islands [9]. Interestingly, many co-occurring species significantly differ in mouthpart sizes and shapes, presumably owing to adaptations to a specialized diet [6,7], suggesting that stenophagy has evolved multiple times independently in these islands [10]. Although behavioral and physiological experiments have revealed a close correlation between morphological traits and prey preference in *Dysdera*, little is known about the molecular basis of trophic adaptations in this genus.

Here we present the draft assembly and functional annotation of the genome of the Canary Island endemic spider *Dysdera silvatica* Schmidt, 1981 (NCBI:txid477319; Fig. 1). This study is the first genomic initiative within its family and just the second within the Synspermiata [11], a clade that includes most of the families formerly included in Haplogynae, which was recently shown to be paraphyletic [12,13] (Fig. 2). Remarkably, a recent review on arachnid genomics identified the superfamily Dysderoidea (namely, Dysderidae, Orsolobidae, Oonopidae, and Segestriidae) as one of the priority candidates for genome sequencing [14]. The new genome, intended to be a reference genome for genomic studies on trophic specialization, will also be a valuable source for the ongoing studies on the molecular components of the chemosensory system in chelicerates [15]. Besides, because of the numerous instances of independent adaptation to caves [16], the peculiar holocentric chromosomes [17], and the evidence for cryptic female choice mechanisms [18,19] within the family, the new genome will be a useful reference for the study of the molecular basis of adaptation to extreme environments, karyotype evolution, and sexual selection. Additionally, a new fully annotated spider genome will greatly improve our understanding of key features, such as the venom and silk. The availability of new genomic information in a sparsely sampled section of the tree of life of spiders [14] will further provide valuable knowledge about relevant scientific questions, such as gene content evolution across main arthropod groups, including the consequences of whole-genome duplications, or the phylogenetic relationships with Araneae.

Sampling and DNA extraction

We sampled adult individuals of *D. silvatica* in different localities of La Gomera (Canary Islands) in March 2012 and June 2013 (Supplementary Table S1-1). The species was confirmed in the laboratory, and samples were stored at −80°C until its use. For Illumina and PaciBio libraries (see below), we extracted genomic DNA using Qiagen DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany, 74104)) according to the manufacturer's protocol. For the Oxford Nanopore libraries, we used a modified version of the Blood & Cell Culture DNA Mini Kit (Qiagen). Due to the high amount of chitin present in spiders we incubated fresh original samples 48 h at 32°C, avoiding a centrifugation step prior to sample loading to Qiagen Genomic tips, permitting the solution to precipitate by gravity. We also added an extra wash with 70% ethanol and centrifuged the solution at >5,000*g* for 10 min at 4°C. We quantified the genomic DNA in a Qubit fluorometer (Life Technologies, Thermo Fisher Scientific Inc., USA) using the dsDNA BR (double stranded DNA Broad Range) Assay Kit and checked its purity in a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific Inc.).

DNA sequencing

We sequenced the genome of *D. silvatica* using 4 different sequencing platforms (Table 1; Supplementary Table S1-2). First, we used the Illumina HiSeq2000 to obtain the genome sequence of a single male (100 bp, paired-end [PE] reads, 100 PE; TruSeq library). The flow-cell lane generated ∼51 Gb of sequence, representing a genome coverage of 30× (assuming a genome size of ∼1.7 Gb; see below). The genome of a female was sequenced using a mate pair (MP) approach; for that we used Nextera 5 kb-insert 100 PE libraries and the HiSeq2000 to generate ∼40 Gb of sequence (∼23× of coverage). A third individual (male) was used for single-molecule real-time (SMRT) sequencing (PacBio long reads). We used 8 SMRT libraries (20 kb SMRT bell templates), which were sequenced using the P6-C4 chemistry in a PacBio RSII platform. We obtained a yield of ∼9.6 Gb (raw coverage of

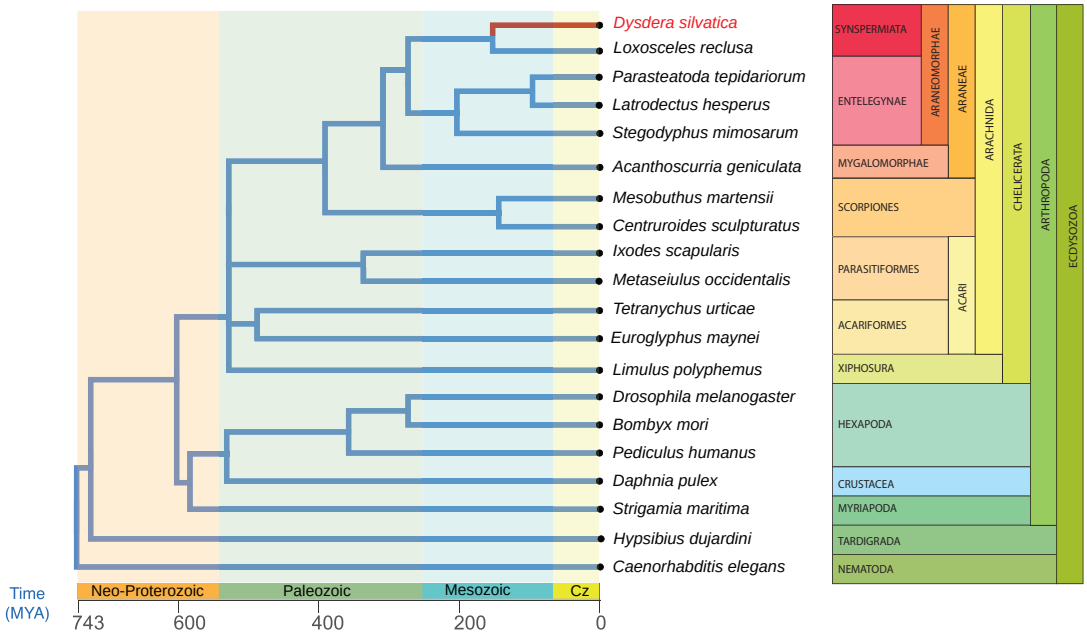


Figure 2 Phylogenetic relationships of the species used for the *D. silvatica* genome annotation (see Supplementary Table S1-11 for further details) and completeness analysis. Because the chelicerata phylogeny is controversial (e.g., [20], [21]), we set the most conflictive clades as polytomies. Divergence times were obtained from Carlson et al. (2017) [22] and the TimeTree web server (http://www.timetree.org/). Cz, cretaceous period.

Table 1. Sequencing data and library information

Run ID	Library	Insert size	Read lengths	Lanes	Total bases	Raw read pairs	Coverage (×)[a]
PE	Illumina HiSeq200 - Truseq	370 bp	100×100 PE	1	51,202,445,102	506,954,902	30
MP	Illumina HiSeq200 - Nextera	5 kb	100×100 PE	1	39,609,522,995	392,173,495	23
Nanopore	Nanopore 1D Libraries	-	Nanopore	5	23,193,357,481	20,534,058	14
PacBio	PacBio RSII 20 Kb SMRTbell	-	SMRT	8	9,652,844,880	1,455,288	6

[a]Based on the genome size estimated by flow cytometry ~1.7 Gb.

~6×). Finally, 2 additional females were used for the 5 runs of Nanopore sequencing (Nanopore 1D libraries). We got a yield of ~23.2 Gb (~14× coverage) (Table 1; Supplementary Table S1-2).

D. silvatica chromosome and genome size

D. silvatica has a diploid chromosome set of 6 pairs of autosomes and 2 (females are XX; 2n = 14) or 1 (males are X0) sex chromosomes (M. A. Arnedo, unpublished results). Using flow cytometry and the genome of the German cockroach *Blattella germanica* (1C = 2.025 Gb, J. S. Johnston, personal communication; see also [23]) as reference, we determined that the haploid genome size of *D. silvatica* is ~1.7 Gb. For the analysis, we adapted the Hare and Johnston [24] protocol for spiders species, without using male palps and chelicers to avoid analyzing haploid or endoreplicated cells, respectively [25,26]. Shortly, we isolated cells from the head of the male cockroach, and legs and palps from female spiders. We incubated the cells in LB0.1 with 2% of tween [27], propidium iodide (50 μg/mL), and RNAse (40 μg/mL). After 10 minutes, the processed tissue was filtered using a nylon mesh of 20 μm. We determined the DNA content of the diploid cells through the relative G0/G1 peak positions of the stained nuclei using a Gallios flow cytometer (Beckman Coulter, Inc, Fullerton, CA); the results were based on the average of 3 spider replicates, counting a minimum of 5,000 cells per individual.

In addition, we also estimated the *D. silvatica* genome size from the distribution of *k*-mers (from short reads) with Jellyfish v.2.2.3 (Jellyfish, RRID:SCR_005491) [28]. The distribution of *k*-mers of size 17, 21, and 41 (GenomeScope (GenomeScope, RRID:SCR_017014) [29]) resulted in a haploid genome size of ~1.23 Gb (Supplementary Fig. S1). The discrepancy between *k*-mer– and cytometry-based estimates may be caused by the presence of repetitive elements [30], which can affect *k*-mer estimates.

Read preprocessing

To avoid including contaminants in the assembly step, we searched the raw reads for mitochondrial, bacterial, archaeal, and virus sequences. We downloaded all genomes of all these kinds available in the GenBank database (Supplementary Table S1-3) and used BLASTN v2.4.0 (BLASTN, RRID:SCR_001598) [31] to detect and filter all contaminant reads (E-value $<10^{-5}$;

>90% alignment length; >90% identity). We preprocessed raw reads using PRINSEQ v.0.20.3 (PRINSEQ, RRID:SCR_005454) [32]. We estimated some descriptive statistics, such as read length and *k*-mer representation, and calculated the amount of adapter sequences and exact duplicates.

Quality-based trimming and filtering was performed according to the chemistry, technology, and library used (Supplementary Table S1-4). For the short-insert 100 PE library, we used Trimmomatic v0.36 (Trimmomatic, RRID:SCR_011848) [33] with specific lists of adapters of the TruSeq v3 libraries to filter all reads shorter than 36 bp or with minimum quality scores < 30 along 4-bp sliding windows. We also filtered trailing and leading bases with a quality score < 10. Long-insert MP libraries were preprocessed using NxTrim v0.4.1 [34] with default parameters (Supplementary Table S1-4a and b). We preprocessed the raw PacBio reads using the SMRT Analysis Software (SMRT Analysis Software, RRID:SCR_002942) [35], by generating circularized consensus sequence to further perform a polishing analysis with Pilon v1.22 (Pilon, RRID:SCR_014731) [36] based on short reads (Supplementary Table S1-4c).

De novo genome assembly

We used MaSuRCA v3.2.9 (MaSuRCA, RRID:SCR_010691) [37] for a hybrid *de novo* assembly of the *D. silvatica* genome (Supplementary Fig. S2). Additionally, we performed a scaffolding phase using AGOUTI (minimum number of joining reads pairs support, *k* = 3) [38], and the raw reads from a *D. silvatica* RNA sequencing (RNAseq) experiment [39] (Supplementary Table S1-5 and S1-6). During the assembly phase, we chose for each software the parameter values that generated the best assembly (Supplementary Table S1-7) in terms of (i) continuity and contig size statistics, such as the N50, L50, and the total number of sequences and bases assembled; and (ii) completeness measures, obtained as the fraction (and length) of a series of highly conserved proteins present in the draft genome. Particularly, we used 5 datasets, BUSCO v3 (BUSCO, RRID:SCR_015008) with genome option [40] using (i) the Arthropoda or (ii) the Metazoa dataset, (iii) the 457 core eukaryotic genes (CEGs) of *Drosophila melanogaster* [41], (iv) the 58,966 transcripts in the *D. silvatica* transcriptome [39], and (v) the 9,473 1:1 orthologs across 5 *Dysdera* species, *D. silvatica*; *D. gomerensis* Strand, 1911; *D. verneaui* Simon, 1883; *D. tilosensis* Wunderlich, 1992; and *D. bandamae* Schmidt, 1973 obtained from the comparative transcriptomics analysis of these species [42]. Finally, we performed an additional search to identify and remove possible contaminants in the generated scaffolds (Supplementary Table S1-7). We discarded 16 contaminant sequences > 5 kb. The final assembly size of the *D. silvatica* genome (Dsil v1.2) was ~1.36 Gb, with an N50 of ~38 kb (Table 2).

We determined the average genome coverage for each sequencing library with SAMtools v1.3.1 (SAMtools, RRID:SCR_002105) [43], by mapping short reads (using bowtie2 v2.2.9 [bowtie2, RRID:SCR_005476] [44]) or long reads (using minimap2 [45]) to the final draft assembly (Table 1; Supplementary Table S1-8; Supplementary Fig. S3).

Repetitive DNA sequences

We analyzed the distribution of repetitive sequences in the genome of *D. silvatica*, using either a *de novo* with RepeatModeler v1.0.11 (RepeatModeler, RRID:SCR_015027) [46], or a database-guided search strategy with RepeatMasker v.4.0.7 (RepeatMasker, RRID:SCR_012954) [47]. We used 3 different databases of repetitive sequences, (i) *D. silvatica*–specific repetitive elements generated with RepeatModeler v1.0.11 [46], (ii) the Dfam_Consensus [48] (version 20170127), and (iii) the RepBase (version 20170127) [49,50]. We identified 2,604 families of repetitive elements, where 1,629 of them (62.6%) were completely unknown. Repetitive sequences accounted for ~732 Mb, which represent 53.8% of the total assembly size (Table 2; Supplementary Table S1-9a). Remarkably, most abundant repeats are from unknown families, 22.6% of the assembled genome. The repetitive fraction of the genome also include DNA elements (16.8%), LINEs (10.7%), and SINEs (1.85%), and a small fraction of other elements, including LTR elements, satellites, simple repeats, and low-complexity sequences. We found that the 10 most abundant repeat families among the 2,604 identified in *D. silvatica* account for ~7% of the genome and encode 5 unknown, 3 SINEs, and 2 LINEs, with an average length of ~193, ~161, and ~1,040 bp, respectively (Supplementary Table S1-9b).

We also studied the distribution of the high-covered genome regions to describe the spacing pattern among repetitive sequences. In particular, we searched for genomic regions that have a higher than average sequencing coverage above a particular threshold. Because repetitive regions are more prone to form chimeric contigs in the assembly step, we only used MaSuRCA super reads, and longer than 10 kb and free of Ns (34,937 contigs; 1.12 Gb). We estimated the coverage after mapping the short reads (from the 100PE library) to those contigs. We defined as high-coverage regions (HCRs) those with a coverage ≥2.5× or 5× the genome-wide average (~30×), in a region of ≥150, ≥500, ≥1,000, or ≥5,000 bp (Supplementary Fig. S4a; Supplementary Table S2). We found a large number of contigs encompassing ≥1 HCR. For instance, 21,614 contigs (~61.9%) include ≥1 HCR of 150 bp with >2.5× coverage (an average of 2.48 HCRs per contig; 77.7 HCR per Mb) (Supplementary Table S2-2a). For HCRs of >5× coverage, the results are also remarkable (10,604 contigs have ≥1 HCR of 150 bp, corresponding to 25.6 HCR per Mb). As expected, the longer the HCR the smaller the fraction in the genome; indeed, we found that the genome is encompassing ~5 HCR per Mb (HCR, longer than 1 kb at 2.5×). The distances between consecutive HCRs doenot show clear differences between the 2.5× and 5× thresholds (Supplementary Fig. 4b and S5; Supplementary Table S2-2b).

Table 2. *Dysdera silvatica* nuclear genome assembly and annotation statistics

Genome assembly[a]	Value
Assembly size (bp)	1,359,336,805
% AT/CG/N	64.91%/34.83%/0.26%
Number of scaffolds	65,205
Longest scaffold	340,047
N50	38,017
L50	10,436
Repeat statistics[b]	
Number of elements	3,284,969
Length (bp) [% Genome]	731,540,381 [53.81%]
Genome annotation[a]	
Protein-coding genes	48,619
Functionally annotated	36,398 (74.86%)
Without functional annotation	12,221 (25.14%)
tRNA genes	33,934

[a]See also Supplementary S1-7.
[b]Summary of the RepeatMasker analysis (See also Supplementary Table S1-9).

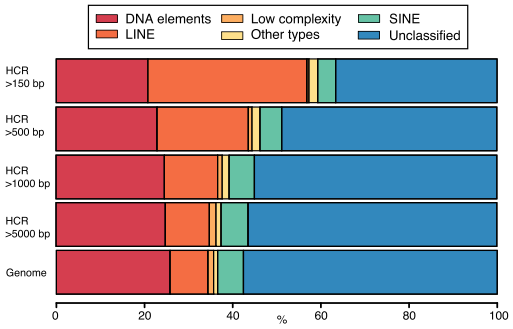


Figure 3 Bar plot of the annotation of the repetitive elements within the HCRs (2.5× threshold) at different intra-HCR length cutoffs (150, 500, 1,000, and 5,000 bp) (Supplementary Table S2-2a). Colors represent the type of repeat element identified by RepeatMasker. "Other types" class includes the LTR elements, small RNA, and satellite information that represent a small fraction.

We found a strong relationship between the length of the HCR and the type of the included repetitive elements (Fig. 3; Supplementary Table S2-3). For instance, while LINEs represent 8.62% of the repetitive elements in the whole genome, they are clearly enriched in the HCRs (36.12% in HCRs longer than 150 bp; 12.08% in HCRs longer than 5,000 bp) (Fig. 3; Supplementary Table S2-3a); the same was found for the small RNA fraction (ribosomal RNA). In contrast, the fraction of low-complexity repetitive sequences is much less represented in small HCRs than in the whole genome (~1.3%). We also found that the coverage threshold has little effect on the results (Supplementary Table S2-3; Supplementary Fig. S6), either for the main families or across subfamilies (Supplementary Table S2-4 and S2-5).

Given that the HCR analysis covers an important fraction of the assembled bases (~82%), the present results can likely be extrapolated to the whole genome. Therefore, the relatively low N50 of the *D. silvatica* genome draft is very likely to be caused by abundant interspersed repeats preventing genome continuity. Despite the low N50 we estimated that the draft presented here is mostly complete in terms of functional regions (see below).

Transcriptome assembly and genome annotation

We used the newly generated genome sequence to obtain a reference-guided assembly of the *D. silvatica* transcriptome with the RNAseq data from Vizueta et al. [39]. We used HISAT2 v2.1.0 (HISAT2, RRID:SCR_015530) [51] to map the RNAseq reads to the reference and Trinity v2.4.0. (Trinity, RRID:SCR_013048) [52] (genome-guided bam, max intron = 50 kb, min coverage = 3) to assemble the transcriptome (named "Dsil-RefGuided transcriptome"; Supplementary Table S1-10). We used the MAKER2 v2.31.9 (MAKER2, RRID:SCR_005309) [53] genome annotation pipeline for the structural annotation of *D. silvatica* genes (Supplementary Fig. S2), using both *ab initio* gene predictions and annotation evidences from *D. silvatica* and other sources. For the *ab initio* gene predictions we initially trained Augustus v3.1.0 (Augustus, RRID:SCR_008417) [54] and SNAP (SNAP, RRID:SCR_002127) [55] softwares using scaffolds longer than 20 kb, and BUSCO gene models generated from completeness searches. Then we iteratively included a reliable set of proteins for a further training. This dataset was composed of the 9,473 orthologs 1:1 identified in 5 *Dysdera* species and the 1:1 orthologs among spiders available at OrthoDB v10 (OrthoDB, RRID:SCR_011980) [56] (8,792). After several iterative training rounds, we applied MAKER2, Augustus, and SNAP, adding other sources of evidence: (i) transcript evidence (Dsil-RefGuided transcriptome), (ii) RNAseq reads exon junctions generated with HISAT2 [51] and regtools [57], and (iii) proteins annotated in other arthropods, especially chelicerates (Fig. 2; Supplementary Table S1-11). The annotation process resulted in 48,619 protein-coding and 33,934 transfer RNA (tRNA) genes. The mean annotation edit distance (AED) upon protein-coding genes was 0.32 (Supplementary Fig. S6), which is typical of a well-annotated genome [58, 59]. After each training and iterative annotation round, we checked the improvement of the annotation by means of the cumulative fraction of AED (Supplementary Table S1-12a; Supplementary Fig. S7).

We searched for the presence of protein domain signatures in annotated protein-coding genes using InterProScan v5.15-54 (InterProScan, RRID:SCR_005829) [60,61], which includes information from public databases (see additional details in Supplementary Table S1-7). Additionally, we used NCBI BLASTP v.2.4.0 (BLASTP, RRID:SCR_001010) [31] (E-value cutoff $<10^{-5}$; >75% alignment length) against the Swiss-Prot database to annotate *D. silvatica* genes. We found that 74.9% (36,398 genes) of the predicted protein-coding genes have hits with records of either InterPro (32,322 genes) (InterPro, RRID:SCR_006695) or Swiss-Prot (17,225 cases) (Table 2; Supplementary Table S1-7).

Table 3. Completeness analysis[a]

BLAST analysis[b]	Number Identified (%)
Parasteatoda genes ($n = 30,041$)	19,580 (65.2)
Single-copy *Dysdera* ($n = 9,473$)	8,420 (88.9)
Single-copy spiders ($n = 2,198$)	2,141 (97.4)
CEG ($n = 457$)	438 (95.8)
BUSCO analysis[c]	
Metazoa ($n = 978$)	
Identified BUSCO	882 (90.2)
Complete (C)	689 (70.5)
Single copy (S)	662 (67.7)
Duplicated (D)	27 (2.8)
Fragmented (F)	193 (19.7)
Missing (M)	96 (9.8)
Artrhopoda (**$n = 1,066$**)	
Identified BUSCO	959 (89.9)
Complete (C)	736 (69.1)
Single copy (S)	702 (65.9)
Duplicated (D)	34 (3.2)
Fragmented (F)	223 (20.9)
Missing (M)	107 (10.0)

[a]Completeness analysis of the 36,398 functional annotated proteins of *D. silvatica*.
[b]BLASTP searches against different datasets. E-value cutoff $< 10^{-3}$, alignment length cutoff > 30%, and identity cutoff > 30%.
[c]BUSCO analysis using default parameters against different datasets (BUSCO, RRID:SCR_015008).

Completeness

We determined the completeness of the *D. silvatica* genome assembly (Table 3) using BLASTP (E-value cutoff $<10^{-3}$; >30% of alignment length and identity > 50%). We searched for homologs of the functionally annotated peptides (36,398) (i) among CEG genes of *Drosophila melanogaster* [41]; (ii) among the pre-

dicted peptides of *Parasteatoda tepidariorum*, a spider with a well-annotated genome [62]; (iii) among the 9,473 1:1 orthologs across 5 *Dysdera* species; and (iv) among the 2,198 single-copy genes identified in all spiders and available in OrthoDB v10 [56]. We found in *D. silvatica* a high fraction of putative homologs (95.8% of CEG genes, and 97.4% spider-specific single-copy genes; Table 3). Furthermore, the analysis based on the putative homologs of the single-copy genes included in the BUSCO dataset (BUSCO, RRID:SCR_015008) [40], applying the default parameters for the genome and protein mode, also demonstrated the high completeness of the genome draft. Indeed the analysis recovered the ∼90% of Metazoa or Arthropoda genes (v9), and nearly 70% of them are complete in *D. silvatica*.

We extended the search for *D. silvatica* homologs to a broader taxonomic range (Fig. 2; Supplementary Table S1-11) by including other metazoan lineages and performing a series of local BLASTP searches (E-value cutoff $< 10^{-3}$; >30% alignment length). We found that a great majority of *D. silvatica* genes are shared among arthropods (57.9%), 11,995 of them (32.95%) also being present in Ecdysozoa (Fig. 4a). Remarkably, 9,560 genes appears to be spider-specific, 4,077 of them being specific (unique) of *D. silvatica*. Despite almost all these species-specific genes having interproscan signatures, the annotation metrics are poor compared with genes having homologs in other species (Supplementary Table S1-12b; Supplementary Figs S7 and S9); indeed, they have an average number of exons (2.8) and gene length (∼168aa), which may reflect their partial nature. They could be part of very large genes interspersed by repeats or complex sequences difficult to assemble. The analysis using OrthoDB (v10) [56] across 5 chelicerates (including *D. silvatica*) identified 1,798 genes, with 1:1 orthologous relationships (Fig. 4b), while 12,101 *D. silvatica* genes showed other more complex orthologous/homologous relationships (Fig. 4b, Supplementary Table S1-12c and S3-1). The analysis across the genome annotations of some representative arthropods identified 950 genes with 1:1 orthologous relationships (Supplementary Fig. S8, Supplementary Table S1-12c and S3-2).

Mitochondrial genome assembly and annotation

We assembled the mitochondrial genome of *D. silvatica* (mtDsil) from 126,758 reads identified in the 100PE library by the software NOVOPlasty [63]. Our *de novo* assembly yielded a unique contig of 14,440 bp (coverage of 878×) (Supplementary Table S1-13). CGVIEW (CGVIEW, RRID:SCR_011779) [64] was used to generate a genome visualization of the annotated mtDsil genome (Supplementary Fig. S10). We identified 2 ribosomal RNAs, 13 protein-coding genes, and 15 tRNAs (out of the putative 22 tRNAs). Based on the contig length and the inability of standard automatic annotation algorithms to identify tRNA with missing arms, as reported for spiders [65], the complete set of tRNAs is most likely present for this species.

Conclusion

We have reported the assembly and annotation of the nuclear and mitochondrial genomes of the first representative of the spider superfamily Dysderoidea and the second genome of a Synspermiata, one of the main evolutionary lineages within the "true spiders" (Araneomorphae) and still sparsely sampled at the genomic level [14]. Despite the high coverage and the hybrid assembly strategy, the repetitive nature of the *D. silvatica* genome precluded obtaining a high-continuity draft. The characteristic holocentric chromosomes of Dysderidae [17] may also explain the observed genome fragmentation; indeed, it has been recently shown that genome-wide centromere-specific repeat arrays are interspersed among euchromatin in holocentric plants (Rhynchospora, Cyperceae) [66].

Nevertheless, the completeness and the extensive annotations achieved for this genome, as well as the new reference-guided transcriptome, make this draft an excellent source tool for further functional and evolutionary analyses in this and other related species, including the origin and evolution of relevant spider traits, such as venom and silk. Moreover, the availability of new genomic information in a lineage with remarkable evolutionary features such as recurrent colonizations of the underground environment or complex reproductive anatomies indicative of cryptic female choice, to cite 2 examples, will further provide valuable knowledge about relevant scientific questions, such as the molecular basis of adaptation to extreme habitats or the genetic drivers of sexual selection, along with more general aspects related to gene content across main arthropod groups, the consequences of whole-genome duplications, or phylogenetic relationships with the Araneae. Additionally, because this genus experienced a spectacular adaptive radiation in the Canary Islands, the present genome draft could be useful to further studies investigating the genomic basis of island radiations.

Availability of supporting data and materials

The whole-genome shotgun project has been deposited at DDBJ/ENA/GenBank under accession number QLNU00000000 and project ID PRJNA475203. The version described in this article is version QLNU01000000. This project repository includes raw data, sequencing libraries information, and assemblies of the mitochondrial and nuclear genomes. Other relevant datasets such as annotation, reference-guide assembled transcripts, repeat, and HCR data, as well as other data relevant for the reproducibility of results, are available in the GigaDB dataset [67].

Additional file

File S1. Supplemental Material Summary
SanchezHerrero_Dsilvatica_SupMaterial_Summary.pdf

Availability of supporting source code and requirements

The scripts employed and developed in this project are available under the github repository:
Project name: Genome assembly of *Dysdera silvatica*
Project home page: https://github.com/molevol-ub/Dysdera_silvatica_genome
Operating system(s): Platform independent
Programming language: Bash, Perl, Python, R
License: MIT

Abbreviations

AED: annotation edit distance; AGOUTI: Annotated Genome Optimization Using Transcriptome Information; BLAST: Basic Local Alignment Tool; bp: base pair; BUSCO: Benchmarking Universal Single Copy Orthologs; CEG: core eukaryotic gene; Cz: Cretaceous period; Dsil: *Dysdera silvatica*; Gb: gigabase pairs; GC: guanine cytosine; GO: Gene Ontology; HCR: high-coverage re-

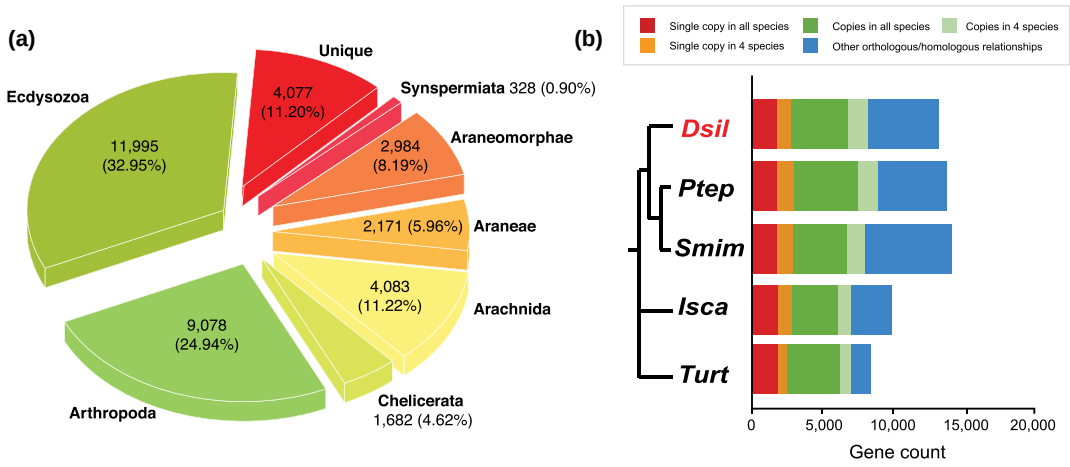


Figure 4 (a) Pie chart illustrating the taxonomic distribution of positive BLAST hits of the *D. silvatica* protein-coding genes against the sequence data of species included in Fig. 2. **(b)** Homology relationships among *D. silvatica* (Dsil) and different chelicerates genomes available in OrthoDB v10 [56], *Parasteatoda tepidariorum* (Ptep), *Stegodyphus mimosarum* (Smim), *Ixodes scapularis* (Isca), and *Tetranychus urticae* (Turt). Red and orange bars indicate the fraction of single-copy genes (1:1 orthologs) identified in all species, and in all but 1 (e.g., missing in 1 species), respectively. The dark and light green bar indicate the fraction of orthologs present in all species and in all but 1, respectively, that are not included in previous categories. The blue bar (other orthology/homology) shows other more complex homologous relationships. The results were generated by uploading *D. silvatica* proteins to the OrthoDB web server.

gions; Isca: *Ixodes scapularis*; kb: kilobase pairs; LINE: long interspersed nuclear element; LTR: long terminal repeats; MaSuRCA: Maryland Super-Read Celera Assembler; Mb: megabase pairs; MP: mate pair; Mya: million years ago; NCBI: National Center for Biotechnology Information; PacBio: Pacific Biosciences; PE: paired-end; PRINSEQ: PReprocessing and INformation of SEQuence data; Ptep: *Parasteatoda tepidariorum*; RNAseq: RNA sequencing; SINE: short interspersed nuclear element; Smim: *Stegodyphus mimosarum*; SMRT: Single-Molecule Real Time; tRNA: transfer RNA; Turt: *Tetranychus urticae*.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by the Ministerio de Economía y Competitividad of Spain (CGL2012-36863, CGL2013-45211, and CGL2016-75255), and by the Comissió Interdepartamental de Recerca I Innovació Tecnològica of Catalonia, Spain (2014SGR-1055 and 2014SGR1604). J.F.S.-H. was supported by a Formación del Profesor Universitario (FPU) grant (Ministerio de Educación of Spain, FPU13/0206); C.F.-L. by an IRBio PhD grant; S.H-A by Becas Postdoctorales en el Extranjero CONACyT; A.S.-G. by a Beatriu de Pinós grant (Generalitat de Catalunya, 2010-BP-B 00175); and J.R. and M.A.A. were partially supported by ICREA Academia (Generalitat de Catalunya).

Authors' contributions

J.R., A.S.-G., and M.A.A designed the study. C.F.-L., J.F.S.-H., P.E., and S.H-A. processed the samples and extracted DNA. J.F.S.-H. performed the bioinformatics analysis and drafted the manuscript. J.F.S.-H., A.S.-G., and J.R. interpreted the data. All authors revised and approved the final manuscript.

Acknowledgments

We acknowledge the Garajonay National Parks for granting collection permits and helping with lodging and logistics during fieldwork. We also thank CNAG (Centro Nacional de Análisis Genómico) for the Nanopore sequencing facilities.

References

1. World Spider Catalog (2018). 2018. http://wsc.nmbe.ch. Accessed on April 2019.
2. Pekár S, Toft S. Trophic specialisation in a predatory group: the case of prey-specialised spiders (Araneae). Biol Rev 2015;**90**(3):744–61.
3. Hopkin SP, Martin MH. Assimilation of zinc, cadmium, lead, copper, and iron by the spider *Dysdera crocata*, a predator of woodlice. Bull Environ Contam Toxicol 1985;**34**:183–87.
4. Pekár S, Líznarová E, Řezác M. Suitability of woodlice prey for generalist and specialist spider predators: a comparative study. Ecol Entomol 2016;**41**(2):123–30.
5. Toft S, Macías-Hernández N. Metabolic adaptations for isopod specialization in three species of *Dysdera* spiders from the Canary Islands. Physiol Entomol 2017;**42**(2):191–98.
6. Řezáč M, Pekár S. Evidence for woodlice-specialization in *Dysdera* spiders: behavioural versus developmental approaches. Physiol Entomol 2007;**32**(4):367–71.
7. Řezác M, Pekár S, Lubin Y. How oniscophagous spiders overcome woodlouse armour. J Zool 2008;**275**(1):64–71.
8. Arnedo MA, Oromí P, Ribera C. Radiation of the spider genus *Dysdera* (Araneae, Dysderidae) in the Canary Islands: cladistic assessment based on multiple data sets. Cladistics 2001;**17**:313–353.
9. Macias-Hernandez N, de la Cruz López S, Roca-Cusachs M, et al. A geographical distribution database of the genus *Dysdera* in the Canary Islands (Araneae, Dysderidae). Zookeys 2016;**625**(625):11–23.
10. Arnedo MA, Oromí P, Múrria C, et al. The dark side of an is-

land radiation: systematics and evolution of troglobitic spiders of the genus *Dysdera* Latreille (Araneae: Dysderidae) in the Canary Islands. Invertebr Syst 2007;**21**(6):623.
11. Michalik P, Ramírez MJ. Evolutionary morphology of the male reproductive system, spermatozoa and seminal fluid of spiders (Araneae, Arachnida) - current knowledge and future directions. Arthropod Struct Dev 2014;**43**(4):291–322.
12. Wheeler WC, Coddington JA, Crowley LM, et al. The spider tree of life: phylogeny of Araneae based on target-gene analyses from an extensive taxon sampling. Cladistics 2017;**33**(6):574–616.
13. Fernández R, Kallal RJ, Dimitrov D, et al. Phylogenomics, diversification dynamics, and comparative transcriptomics across the spider tree of life. Curr Biol 2018;**28**(9):1489–97.
14. Garb JE, Sharma PP, Ayoub NA. Recent progress and prospects for advancing arachnid genomics. Curr Opin Insect Sci 2018;**25**:51–7.
15. Vizueta J, Rozas J, Sánchez-Gracia A. Comparative genomics reveals thousands of novel chemosensory genes and massive changes in chemoreceptor repertories across chelicerates. Genome Biol Evol 2018;**10**(5):1221–36.
16. Deeleman-Reinhold CL. The genus *Rhode* and the harpacteine genera *Stalagtia, Folkia, Minotauria*, and *Kaemis* (Araneae, Dysderidae) of Yugoslavia and Crete, with remarks on the genus *Harpactea*. Rev Arachnol 1993;**10**(6):105–35.
17. Diaz MO, Maynard R, Brum-Zorrilla N. Diffuse centromere and chromosome polymorphism in haplogyne spiders of the families dysderidae and segestriidae. Cytogenet Genome Res 2010;**128**(1-3):131–8.
18. Uhl G. Two distinctly different sperm storage organs in female *Dysdera erythrina* (Araneae: Dysderidae). Arthropod Struct Dev 2000;**29**(2):163–9.
19. Burger M, Kropf C. Genital morphology of the haplogyne spider *Harpactea lepida* (Arachnida, Araneae, Dysderidae). Zoomorphology 2007;**126**(1):45–52.
20. Ballesteros JA, Sharma PP. A critical appraisal of the placement of Xiphosura (chelicerata) with account of known sources of phylogenetic error. Syst Biol 2019, doi:10.1093/sysbio/syz011.
21. Lozano-Fernandez J, Tanner AR, Giacomelli M, et al. Increasing species sampling in chelicerate genomic-scale datasets provides support for monophyly of Acari and Arachnida. Nat Commun 2019;**10**:2295.
22. Carlson DE, Hedin M. Comparative transcriptomics of Entelegyne spiders (Araneae, Entelegynae), with emphasis on molecular evolution of orphan genes. PLoS One 2017;**12**(4):e0174102.
23. Gregory TR. Animal Genome Size Database. 2018. http://www.genomesize.com.
24. Hare EE, Johnston JS. Genome size determination using flow cytometry of propidium iodide-stained nuclei. Methods Mol Biol 2011;**772**:3–12.
25. Rasch EM, Connelly BA. Genome size and endonuclear DNA replication in spiders. J Morphol 2005;**265**(2):209–14.
26. Gregory TR, Shorthouse DP. Genome sizes of spiders. J Hered 2003;**94**(4):285–90.
27. Dpooležel J, Binarová P, Lcretti S. Analysis of nuclear DNA content in plant cells by flow cytometry. Biol Plant 1989;**31**(2):113–20.
28. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 2011;**27**(6):764–70.
29. Vurture GW, Sedlazeck FJ, Nattestad M, et al. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics 2017;**33**(14):2202–4.
30. Austin CM, Tan MH, Harrisson KA, et al. De novo genome assembly and annotation of Australia's largest freshwater fish, the Murray cod (*Maccullochella peelii*), from Illumina and Nanopore sequencing read. GigaScience 2017;**6**(8):1–6.
31. Altschul SF, Gish W, Miller W, et al. Basic Local Alignment Search Tool. J Mol Biol 1990;**215**(3):403–10.
32. Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One 2011;**6**(3):e17288.
33. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;**30**(15):2114–20.
34. O'Connell J, Schulz-Trieglaff O, Carlson E, et al. NxTrim: optimized trimming of Illumina mate pair reads. Bioinformatics 2015;**31**(12):2035–7.
35. PacBio. Single Molecule Real Time (SMRT). https://www.pacb.com/products-and-services/analytical-software/smrt-analysis/.
36. Walker BJ, Abeel T, Shea T, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 2014;**9**(11):e112963.
37. Zimin AV, Marçais G, Puiu D, et al. The MaSuRCA genome assembler. Bioinformatics 2013;**29**(21):2669–77.
38. Zhang SV, Zhuo L, Hahn MW. AGOUTI: improving genome assembly and annotation using transcriptome data. GigaScience 2016;**5**(1):31.
39. Vizueta J, Frias-López C, Macías-Hernández N, et al. Evolution of chemosensory gene families in arthropods: insight from the first inclusive comparative transcriptome analysis across spider appendages. Genome Biol Evol 2017;**9**(1):178–96.
40. Simão FA, Waterhouse RM, Ioannidis P, et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;**31**(19):3210–2.
41. Parra G, Bradnam K, Ning Z, et al. Assessing the gene space in draft genomes. Nucleic Acids Res 2009;**37**(1):289–97.
42. Vizueta, J., Macías-Hernández, N., Arnedo, MA., Rozas, J.and Sánchez-Gracia, A.(2019) Chance and predictability in evolution: the genomic basis of convergent dietary specializations in an adaptive radiation. Mol. Ecol. doi:10.1111/mec.1519931359512
43. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;**25**(16):2078–9.
44. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;**9**(4):357–9.
45. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018;**34**(18):3094–100.
46. Smit AF, Hubley R. RepeatModeler Open-1.0. 2008–2015. http://www.repeatmasker.org.
47. Smit AF, Hubley R, Green P. RepeatMasker Open-3.0. 1996–2010. http://www.repeatmasker.org.
48. Wheeler TJ, Clements J, Eddy SR, et al. Dfam: a database of repetitive DNA based on profile hidden Markov models. Nucleic Acids Res 2012;**41**(D1):D70–D82.
49. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 2015;**6**(1):11.
50. Jurka J, Kapitonov VV, Pavlicek A, et al. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 2005;**110**(1-4):462–7.
51. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods

2015;**12**(4):357–60.
52. Haas BJ, Papanicolaou A, Yassour M, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 2013;**8**(8):1494–512.
53. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 2011;**12**(1):491.
54. Stanke M, Steinkamp R, Waack S, et al. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res 2004;**32**(Web Server issue):W309–12.
55. Korf I. Gene finding in novel genomes. BMC Bioinformatics 2004;**5**(1):59.
56. Kriventseva EV, Kuznetsov D, Tegenfeldt F, et al. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res 2019;**47**(D1):D807–D811.
57. Feng YY, Ramu A, Cotto KC, et al. RegTools: integrated analysis of genomic and transcriptomic data for discovery of splicing variants in cancer. bioRxiv 2018, doi:10.1101/436634.
58. Eilbeck K, Moore B, Holt C, et al. Quantitative measures for the management and comparison of annotated genomes. BMC Bioinformatics 2009;**10**(1):67.
59. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nature Rev Genet 2012;**13**(5):329–42.
60. Mitchell AL, , Attwood TK, Babbitt PC, et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res 2019;**47**(D1):D351–60.
61. Jones P, Binns D, Chang HY, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics 2014;**30**(9):1236–40.
62. Schwager EE, Sharma PP, Clarke T, et al. The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. BMC Biol 2017;**15**(1):62.
63. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res 2016;**45**(4):gkw955.
64. Stothard P, Wishart DS. Circular genome visualization and exploration using CGView. Bioinformatics 2005;**21**(4):537–9.
65. Masta SE, Boore JL. The complete mitochondrial genome sequence of the spider *Habronattus oregonensis* reveals rearranged and extremely truncated tRNAs. Molec Biol Evol 2004;**21**(5):893–902.
66. Marques A, Ribeiro T, Neumann P, et al. Holocentromeres in *Rhynchospora* are associated with genome-wide centromere-specific repeat arrays interspersed among euchromatin. Proc Natl Acad Sci U S A 2015;**112**(44):13633–8.
67. Sánchez-Herrero JF, Frías-López C, Escuer P, et al. Supporting data for "The draft genome sequence of the spider *Dysdera silvatica* (Araneae, Dysderidae): a valuable resource for functional and evolutionary genomic studies in chelicerates." GigaScience Database 2019; http://dx.doi.org/10.5524/100628.

The draft genome sequence of the spider *Dysdera silvatica* (Araneae, Dysderidae): A valuable resource for functional and evolutionary genomic studies in chelicerates

José Francisco Sánchez-Herrero[1,2], Cristina Frías-López[1,2], Paula Escuer[1,2], Silvia Hinojosa-Alvarez[1,2,3], Miquel A. Arnedo[1,4], Alejandro Sánchez-Gracia[1,2,*] and Julio Rozas[1,2,*]

[1]Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona (UB), Barcelona, Spain
[2]Institut de Recerca de la Biodiversitat (IRBio) (UB)
[3]Jardín Botánico, Instituto de Biología, Universidad Nacional Autónoma de México, Ciudad de México, México
[4]Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals (UB)

Additional Files

SUPLEMENTARY FIGURES

Supplementary Figure 1: GenomeScope *k-mer* profile plot for the *D. silvatica* genome *Dsil v1.2*, based on 21-*mers* of the PE reads. The observed *k-mer* frequency distribution is depicted in blue, whereas the GenomeScope fit model is shown as a black line. The unique and putative error k-*mer* distributions are plotted in yellow and red, respectively.

Supplementary Figure 2: Schematic representation of the hierarchical workflow used to generate the assembly of the *D. silvatica* genome.

Supplementary Figure 3: Genome coverage distribution for the different genome sequencing data used in this study. Dash lines indicate the mean genome coverage for the particular sequencing technology.

Supplementary Figure 4: Analysis of the number and distribution of the High Coverage Regions (HCR) across the genome. **a)** Schematic representation of the genome coverage distribution along a contig (~35 kb). The pink dotted line denotes the mean genome coverage estimated for PE read library (Supplementary Table S1-9) (~30X). The green and orange dotted lines reflect, 2.5x and 5x thresholds, respectively, of the average coverage (75X and 150X, respectively). The intra HCR length (in blue) reflects the physical distance fulfilling the threshold coverage (2.5x or 5x), while the Inter HCR (red) denotes the distance between HCRs. **b)** Frequency distribution of the intra-HCR length (blue) and inter-HCR (red) across the 34 937 contigs for the 2.5x (green) or 5x (orange) times the average coverage (See Supplementary Table 2-2 and Supplementary File for details). The minimum value for any inter-HCR was always >10bp. The yellow line denotes the mean distribution value.

Supplementary Figure 5: Frequency distribution of the intra-HCR length (blue) and inter-HCR (red) for different length cutoffs (150, 500, 1 000 and 5 000) across the 34 937 contigs for the 2.5x (a) or 5x (b) threshold coverage (See Supplementary Table 2-2 and Supplementary File for details). The minimum value for any inter-HCR was >10bp. The yellow line denotes the mean distribution value.

Supplementary Figure 6: Bar plot of the annotation of the repetitive elements within the HCRs (5x threshold) at different intra-HCR length cutoffs (150, 500, 1 000 and 5 000 bp) (Supplementary Table S2-2a). Colors represent the type of repeat element identified by RepeatMasker. Other types class, include the LTR elements, Small RNA and Satellites information that represent a small fraction.

Supplementary Figure 7: Cumulative fraction of the frequency distribution of the Annotation Edit Distance (AED) provided by MAKER2 for different steps of the annotation process (Supplementary Table S1-12). The two iterative training rounds (R1 and R2) are shown in dashed blue. The final test (F) rounds are depicted by green lines: F1 using only *D. silvatica* transcripts and F2 using proteins from a broad taxonomic range (Figure 2; Supplementary Table S11). The F2 line shows the cumulative fraction of annotation for the final 48,619 protein-coding genes annotated with an average AED of 0.32. The red and orange dashed lines, represent the cumulative fraction of annotation for the 36,398 functionally annotated protein-coding genes (AED value of 0.268), and for the 4 077 unique *Dysdera silvatica* genes (AED of 0.4), respectively. The AED value is a direct measure of the annotation quality and its values range from 0 (high evidence and exact match based on alignment) to 1 (no evidence support).

Supplementary Figure 8: Homologous relationships between *D. silvatica* (Dsil) and five representative metazoan genomes available in OrthoDB v10 database (Kriventseva 2019): *Strigamia maritima* (Smar), *Drosophila melanogaster* (Dmel), *Limulus polyphemus* (Lpol), *Ixodes scapularis* (Isca), *Parasteatoda tepidariorum* (Ptep) and *D. silvatica* (Dsil). Red and orange bars indicate the fraction of single copy genes (1:1 orthologs) identified in all species, and in all but one (eg, missing in one species), respectively. The dark and light green bar indicates the fraction of orthologs present in all species and in all but one, respectively, that are not included previous categories. The blue bar (other orthology/homology) shows other more complex homologous relationships. The results were generated uploading *D. silvativa* proteins to the OrthoDB web server.

Supplementary Figure S9: Cumulative fraction of frequency distribution of covered exon overlap match by RNAseq or ab initio evidence at the splice site (dash lines) or exon level (solid lines) for the different datasets (in colours: red for species-specific proteins; green for functionally annotated proteins and blue for all structurally annotated proteins) (Supplementary Table S1-12b).

Supplementary Figure S10: Structure and functional annotation of the mitochondrial genome.

SUPLEMENTARY TABLES

Supplementary Table S1-1: Collection of samples used in these study.

Supplementary Table S1-2: DNA sequencing read files used in this study.

Supplementary Table S1-3: NCBI data used for the contaminant search step.

Supplementary Table S1-4: Pre-processing statistics for each library.

Supplementary Table S1-5: Samples used for the RNAseq study.

Supplementary Table S1-6: RNA sequencing read files.

Supplementary Table S1-7: Genome descriptive statistics

Supplementary Table S1-8: Coverage analysis.

Supplementary Table S1-9: RepeatMasker analysis of *D. silvatica* genome.

Supplementary Table S1-10: Reference-guided transcriptome assembly statistics.

Supplementary Table S1-11: Source of proteins to conduct the annotation of *D. silvatica* genes and completeness analysis.

Supplementary Table S1-12: Annotation statistics.

Supplementary Table S1-13: Mitochondrial assembly metrics and annotation features

Supplementary Table S2-1: Example of results for the High Coverage Region (HCR) analysis.

Supplementary Table S2-2: High Coverage Region (HCR) descriptive statistics.

Supplementary Table S2-3: Enrichment analysis of the intersection of High Coverage regions (HCRs) with RepeatMasker annotation (main repeats).

Supplementary Table S2-4: Enrichment analysis of the intersection of High Coverage regions (HCRs) (2.5x mean coverage threshold) with RepeatMasker annotation (subtypes of main repeats).

Supplementary Table S2-5: Enrichment analysis of the intersection of High Coverage regions (HCRs) (5x mean coverage threshold) with RepeatMasker annotation (subtypes of main repeats).

Supplementary Table S3-1: List of the *D. silvatica* genes identified in the OrthoDB analysis across five chelicerates.

Supplementary Table S3-2: List of the *D. silvatica* genes identified in the OrthoDB analysis across six arthropods.

SUPLEMENTARY FILES

Supplementary File SF1: Additional data and results including the mapping coverage distribution results, High Coverage Region (HCR) analysis, Annotation Edit distance (AED) statistics and OrthoDB comparative results.

SUPPLEMENTARY FIGURES

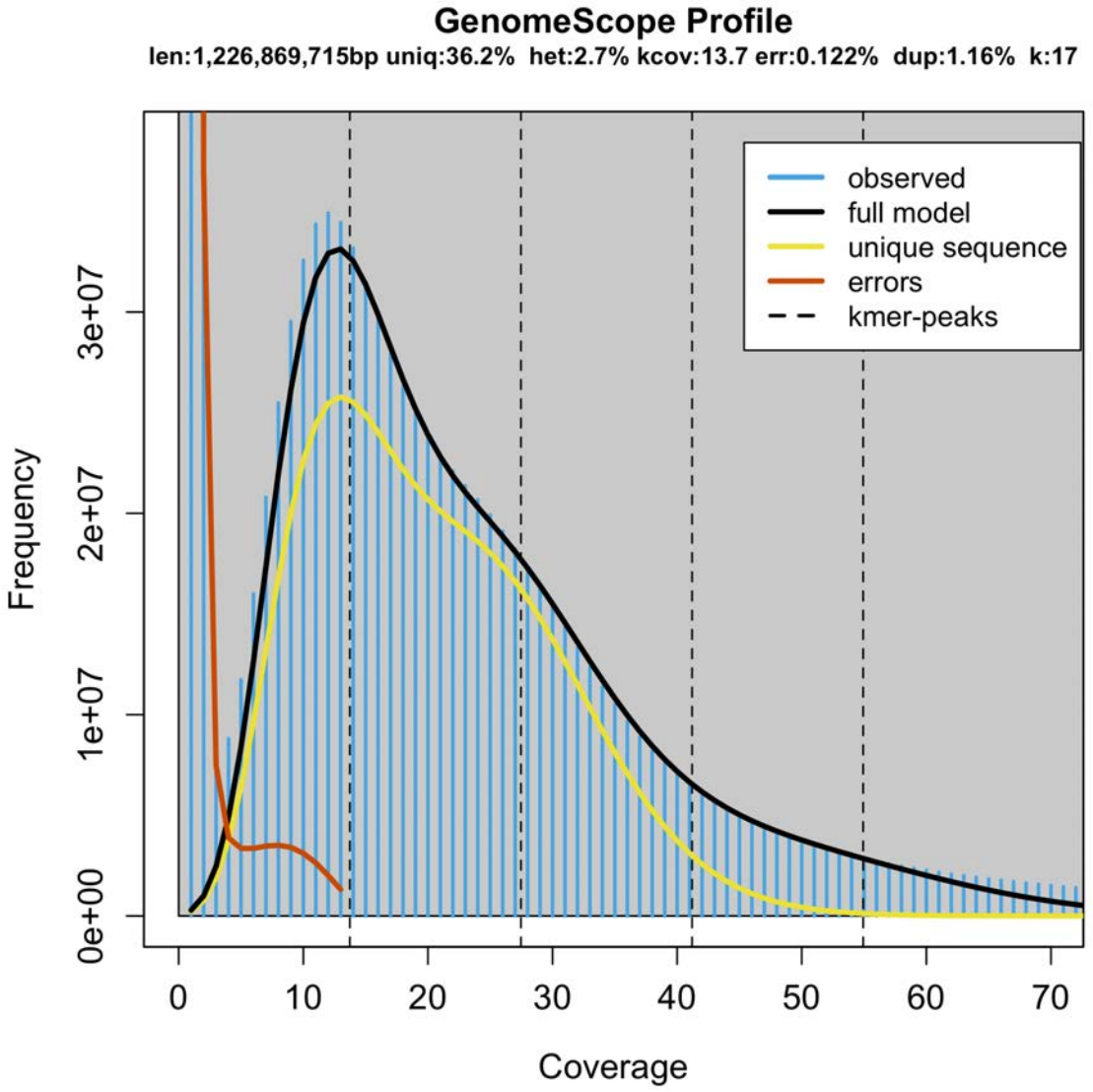

GenomeScope Profile
len:1,226,869,715bp uniq:36.2% het:2.7% kcov:13.7 err:0.122% dup:1.16% k:17
observed
full model
unique sequence
errors
kmer-peaks
Frequency
0e+00
1e+07
2e+07
3e+07
0
10
20
30
40
50
60
70
Coverage

Figure S1

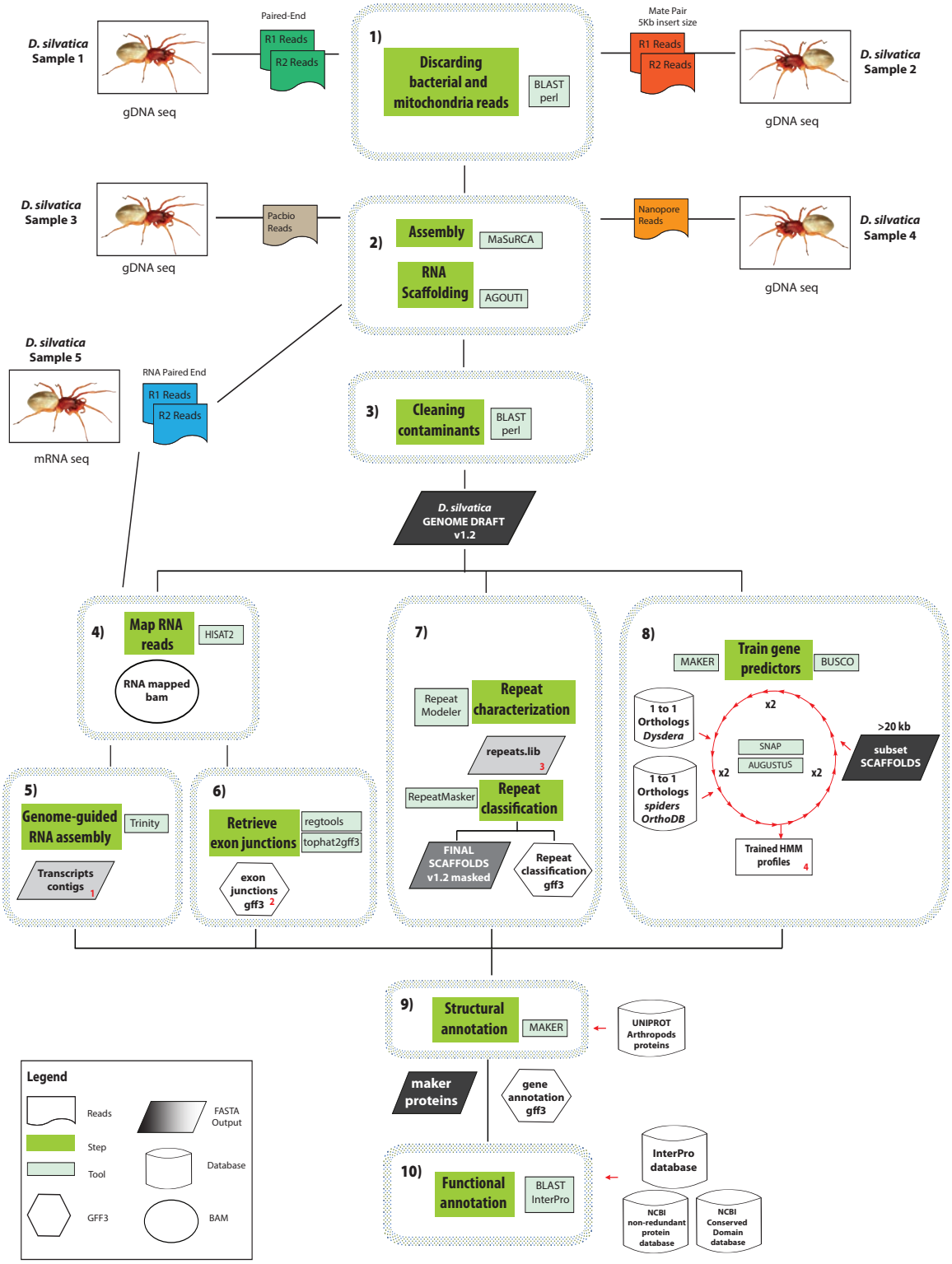

D. silvatica Sample 1
gDNA seq
Paired-End
R1 Reads
R2 Reads
1)
Discarding bacterial and mitochondria reads
BLAST perl
Mate Pair 5Kb insert size
R1 Reads
R2 Reads
D. silvatica Sample 2
gDNA seq
D. silvatica Sample 3
gDNA seq
Pacbio Reads
2)
Assembly
MaSuRCA
RNA Scaffolding
AGOUTI
Nanopore Reads
D. silvatica Sample 4
gDNA seq
D. silvatica Sample 5
mRNA seq
RNA Paired End
R1 Reads
R2 Reads
3)
Cleaning contaminants
BLAST perl
D. silvatica GENOME DRAFT v1.2
4)
Map RNA reads
HISAT2
RNA mapped bam
7)
Repeat Modeler
Repeat characterization
repeats.lib 3
RepeatMasker
Repeat classification
FINAL SCAFFOLDS v1.2 masked
Repeat classification gff3
8)
MAKER
Train gene predictors
BUSCO
1 to 1 Orthologs Dysdera
1 to 1 Orthologs spiders OrthoDB
x2
SNAP
AUGUSTUS
>20 kb
subset SCAFFOLDS
Trained HMM profiles 4
5)
Genome-guided RNA assembly
Trinity
Transcripts contigs 1
6)
Retrieve exon junctions
regtools
tophat2gff3
exon junctions gff3 2
9)
Structural annotation
MAKER
UNIPROT Arthropods proteins
maker proteins
gene annotation gff3
10)
Functional annotation
BLAST InterPro
InterPro database
NCBI non-redundant protein database
NCBI Conserved Domain database
Legend
Reads
Step
Tool
GFF3
FASTA Output
Database
BAM

Figure S2

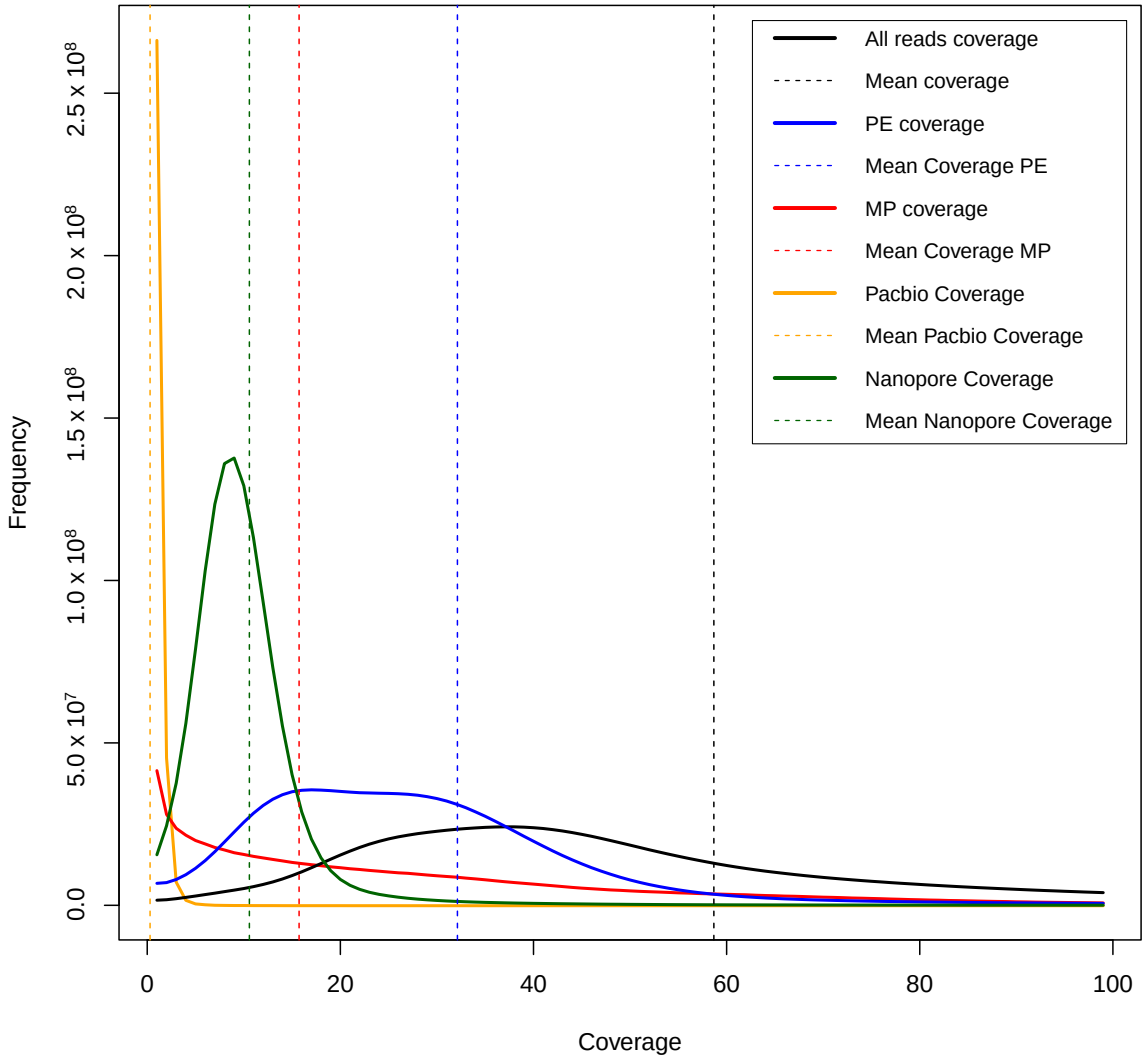

All reads coverage
Mean coverage
PE coverage
Mean Coverage PE
MP coverage
Mean Coverage MP
Pacbio Coverage
Mean Pacbio Coverage
Nanopore Coverage
Mean Nanopore Coverage
Frequency
0.0
5.0 x 10^7
1.0 x 10^8
1.5 x 10^8
2.0 x 10^8
2.5 x 10^8
0
20
40
60
80
100
Coverage

Figure S3

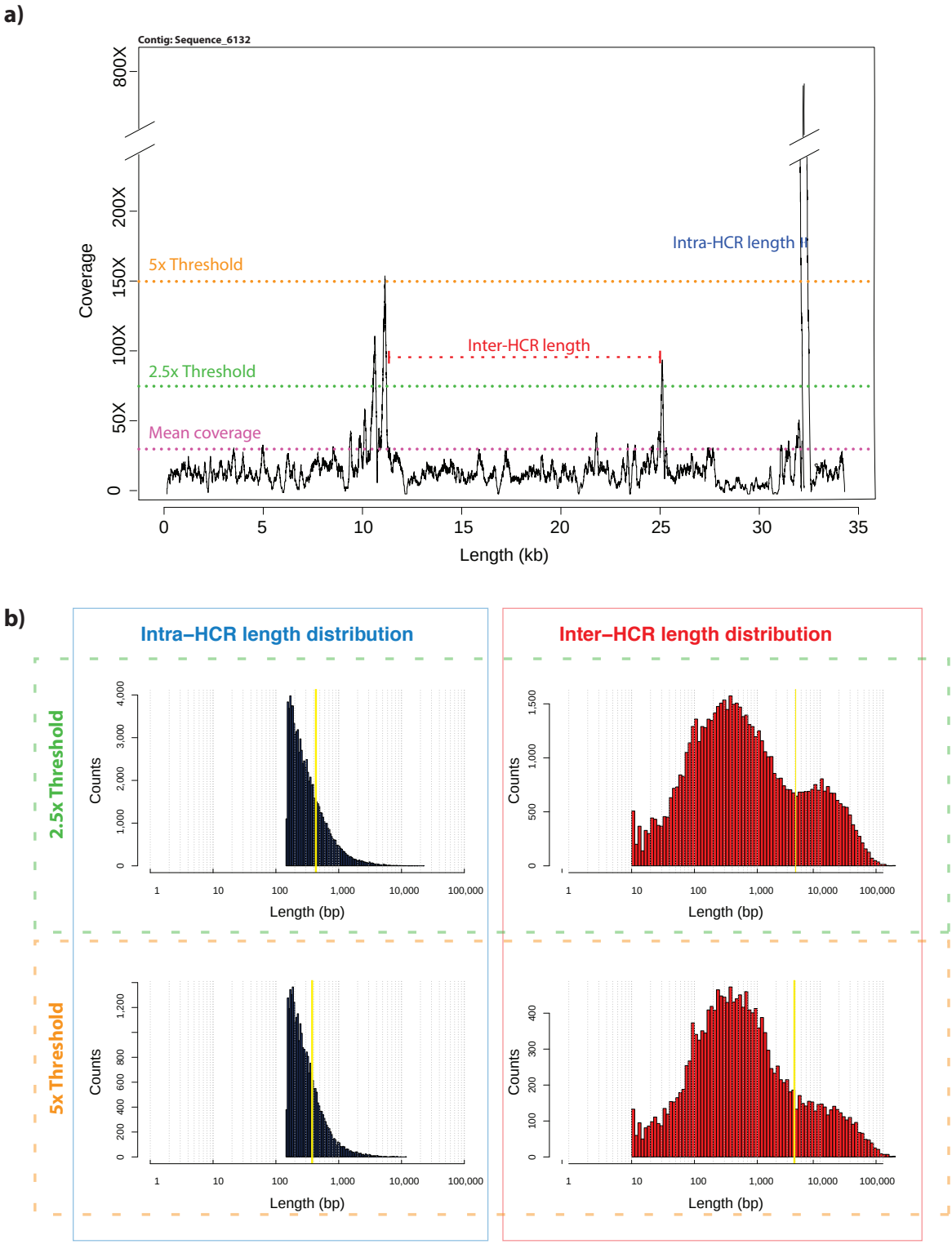

a)
Contig: Sequence_6132
Coverage
800X
200X
150X
100X
50X
0
5x Threshold
2.5x Threshold
Mean coverage
Intra-HCR length
Inter-HCR length
0
5
10
15
20
25
30
35
Length (kb)
b)
Intra–HCR length distribution
Inter–HCR length distribution
2.5x Threshold
5x Threshold
Counts
Length (bp)
1
10
100
1,000
10,000
100,000

Figure S4

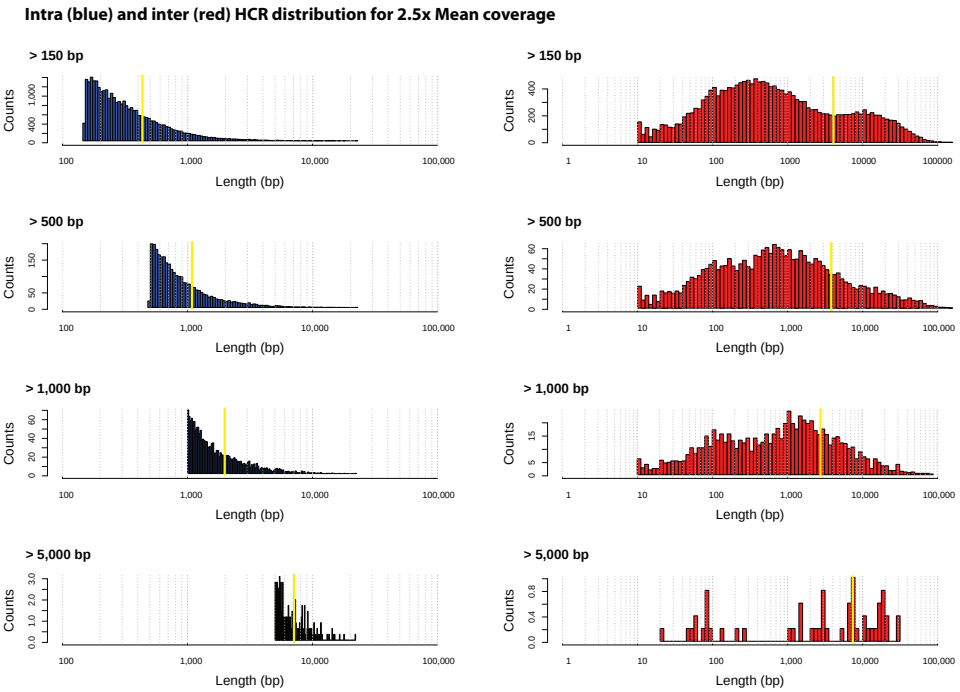


Figure S5a

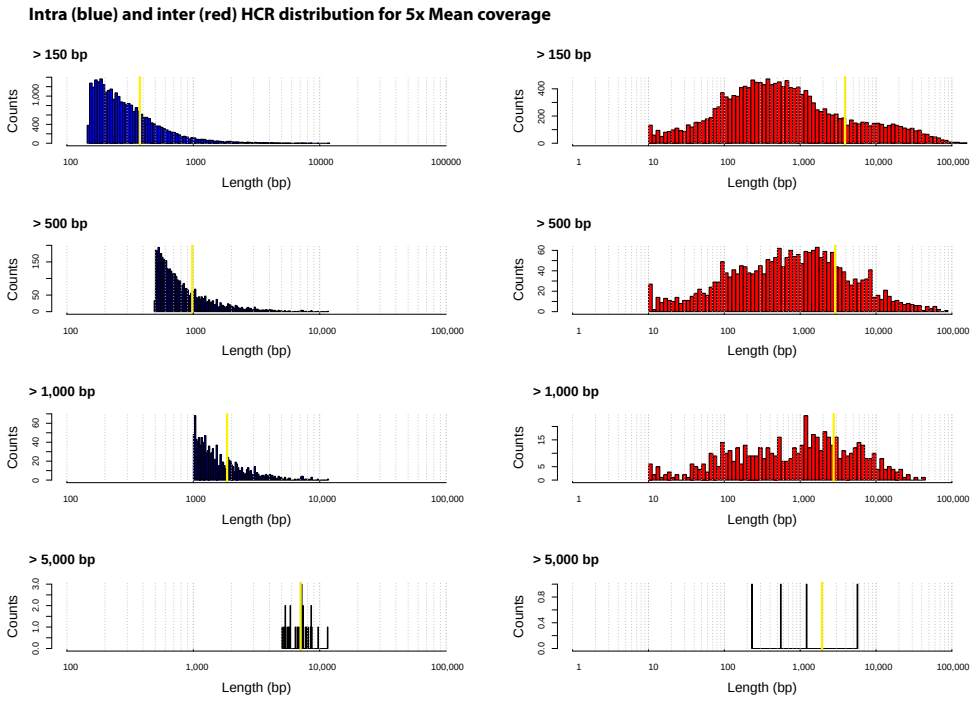


Figure S5b

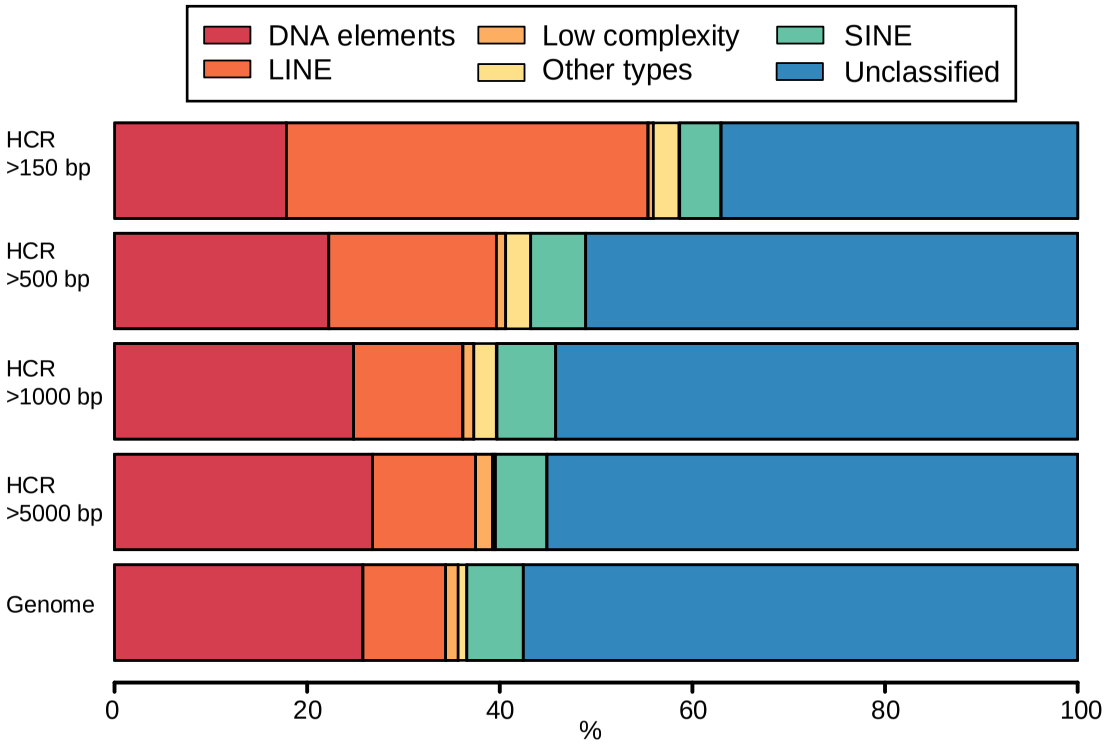

DNA elements
LINE
Low complexity
Other types
SINE
Unclassified
HCR
>150 bp
HCR
>500 bp
HCR
>1000 bp
HCR
>5000 bp
Genome
0
20
40
60
80
100
%

Figure S6

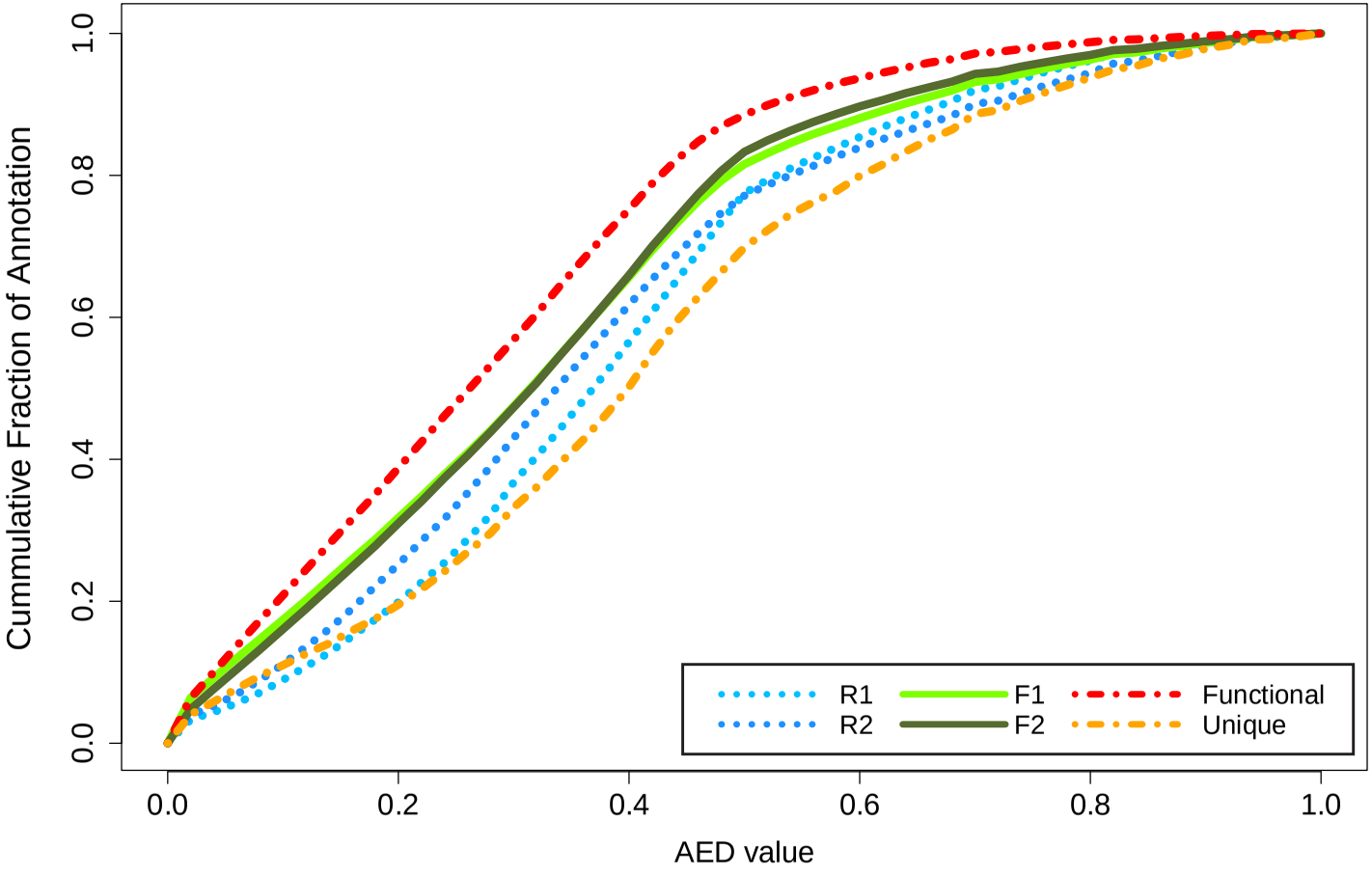

Cummulative Fraction of Annotation
0.0
0.2
0.4
0.6
0.8
1.0
AED value
0.0
0.2
0.4
0.6
0.8
1.0
R1
R2
F1
F2
Functional
Unique

Figure S7

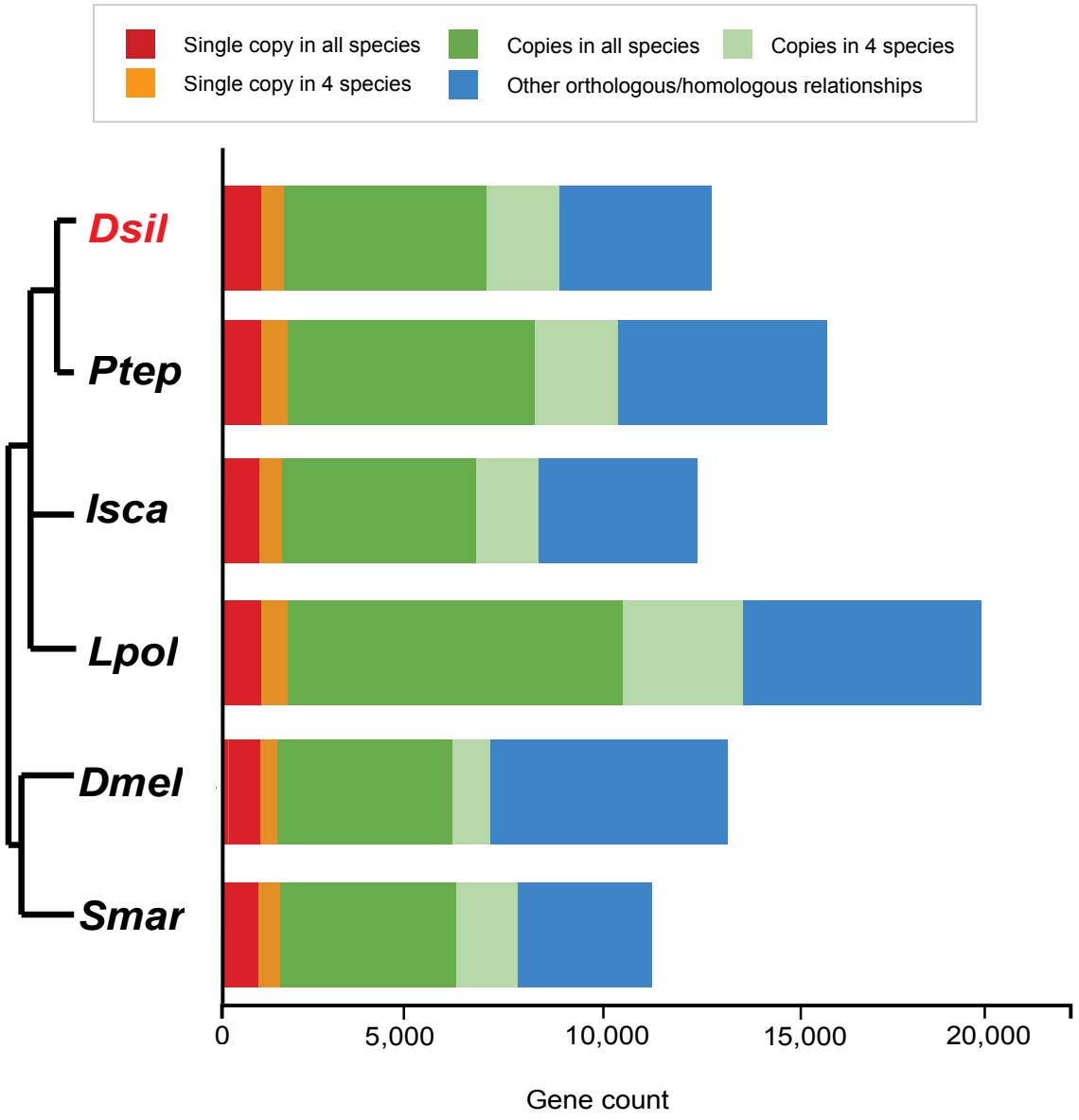

Single copy in all species
Copies in all species
Copies in 4 species
Single copy in 4 species
Other orthologous/homologous relationships
Dsil
Ptep
Isca
Lpol
Dmel
Smar
0
5,000
10,000
15,000
20,000
Gene count

Figure S8

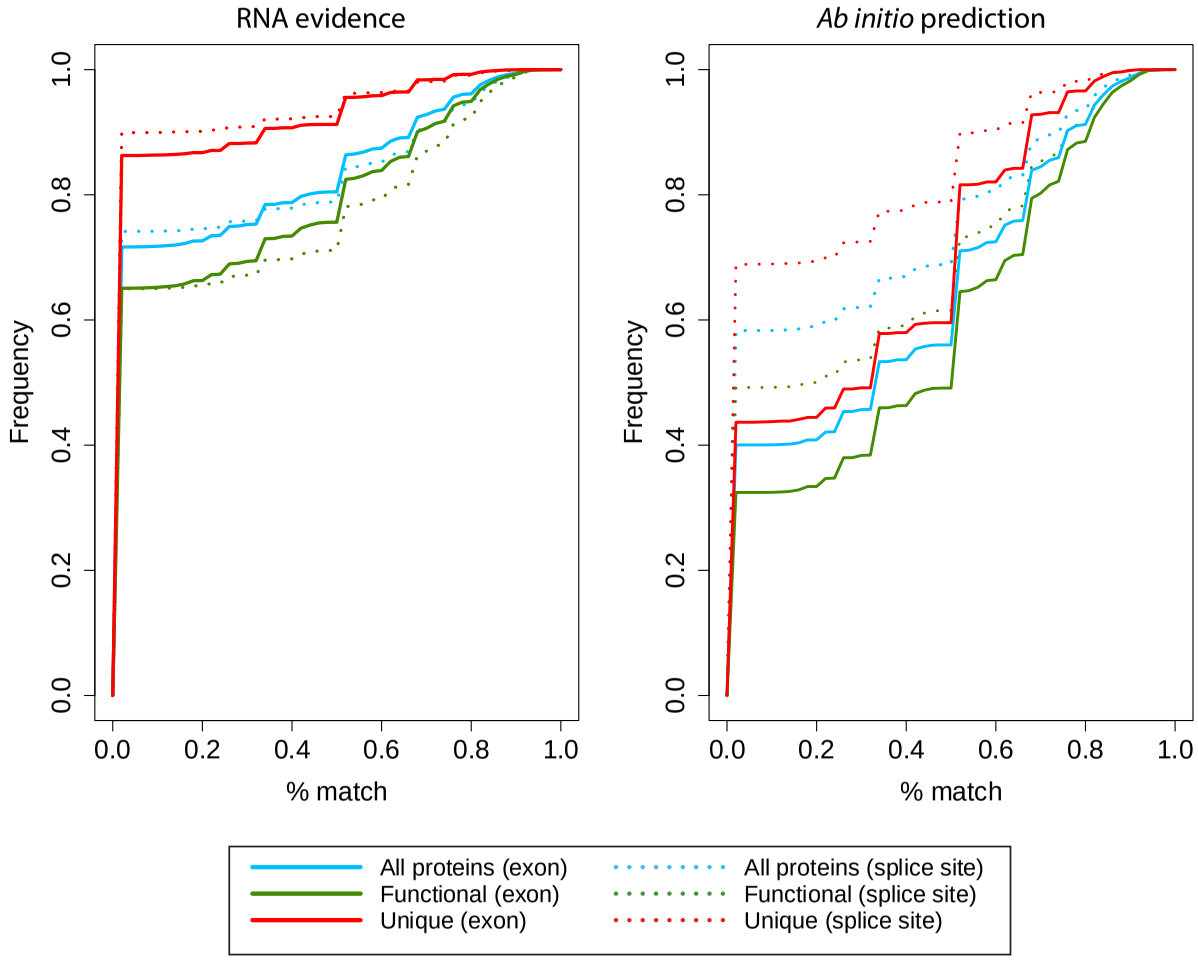

RNA evidence
Ab initio prediction
Frequency
% match
0.0
0.2
0.4
0.6
0.8
1.0
All proteins (exon)
Functional (exon)
Unique (exon)
All proteins (splice site)
Functional (splice site)
Unique (splice site)

Figure S9

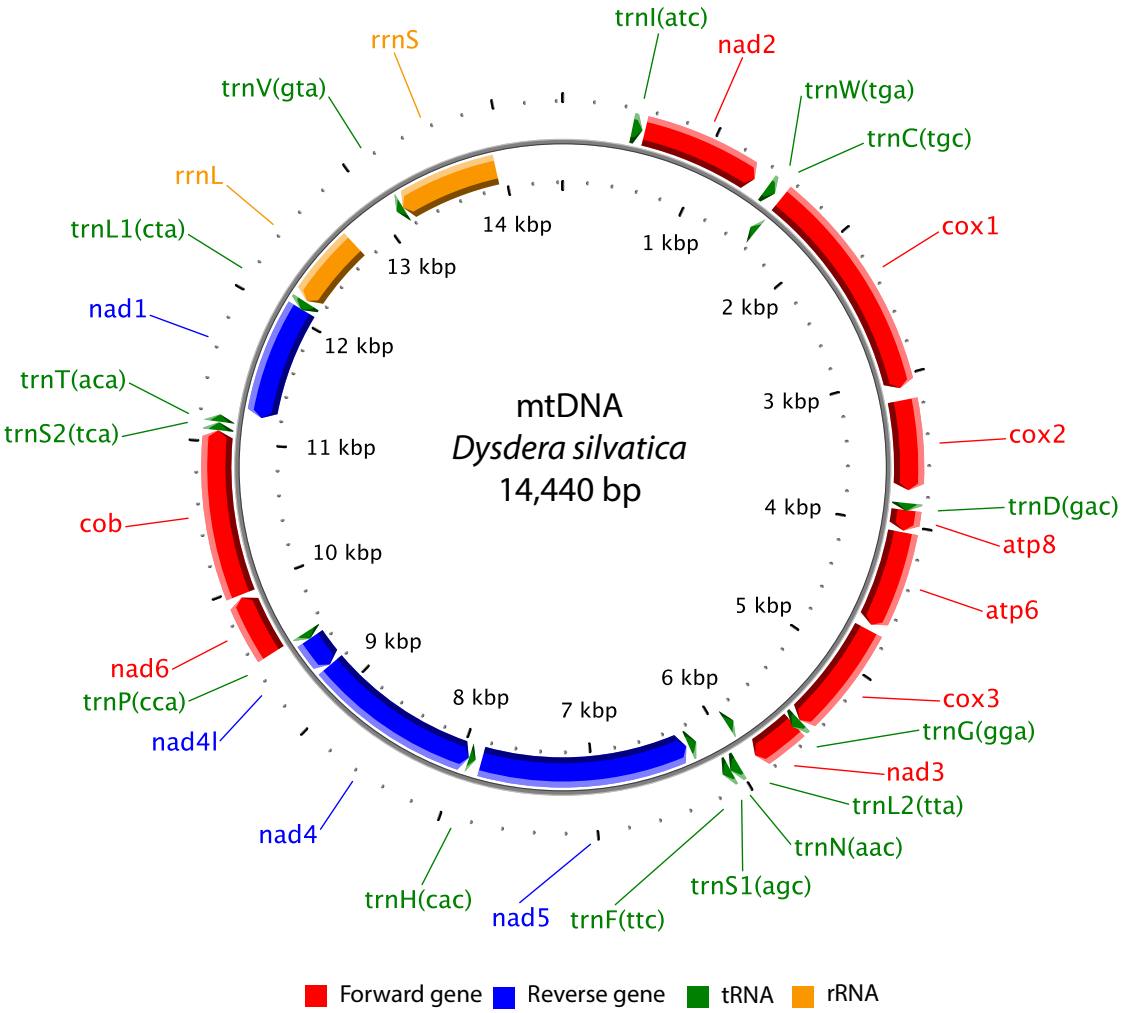

mtDNA
Dysdera silvatica
14,440 bp
1 kbp
2 kbp
3 kbp
4 kbp
5 kbp
6 kbp
7 kbp
8 kbp
9 kbp
10 kbp
11 kbp
12 kbp
13 kbp
14 kbp
trnI(atc)
nad2
trnW(tga)
trnC(tgc)
cox1
cox2
trnD(gac)
atp8
atp6
cox3
trnG(gga)
nad3
trnL2(tta)
trnN(aac)
trnS1(agc)
trnF(ttc)
nad5
trnH(cac)
nad4
nad4l
trnP(cca)
nad6
cob
trnS2(tca)
trnT(aca)
nad1
trnL1(cta)
rrnL
trnV(gta)
rrnS
Forward gene
Reverse gene
tRNA
rRNA

Figure S10

SUPPLEMENTARY TABLES

Table S1-1: **Collection of samples used in these study**

Sample ID	Species	Sex	NCBI BioSample Accession	Sampling Date	Location	Latitude	Longitude	Extraction Protocol	Extraction Date	Application	Run ID	Libraries	Machine	Read Type	Chemistry
NMH2597	*Dysdera silvatica*	Male	SAMN09381544	26-Mar-13	Las Tajoras, La Gomera, Spain	28.112 N	17.262 W	Qiagen DNeasy Blood & Tissue Kit	16-Sep-14	DNA fragment library	Dsilvatica_NMH2597_JFSH-CF_gDNA_PE	Short fragment	Illumina HiSeq 2000	Paired end	TruSeq
NMH2605	*Dysdera silvatica*	Female	SAMN09381545	26-Mar-13	Las Tajoras, La Gomera, Spain	28.112 N	17.262 W	Qiagen DNeasy Blood & Tissue Kit	15-Jan-15	DNA Nextera library	Dsilvatica_NMH2605_JFSH-CF_gDNA_5kb_MP	5 kb Nextera Mate Pair	Illumina HiSeq 2000	Paired end	Nextera
NMH2610	*Dysdera silvatica*	Male	SAMN09389329	04-Jun-13	Teselinde, La Gomera, Spain	28.196 N	17.287 W	Qiagen DNeasy Blood & Tissue Kit	15-May-16	PacBio SMRT libraries	Dsilvatica_NMH2610_JFSH_PacB	20 kb SMRTbell Templates	PacBio RSII	SMRT long reads	P6-C4
NMH1840	*Dysdera silvatica*	Female	SAMN11609372	10-Mar-12	Teselinde, La Gomera, Spain	28.196 N	17.287 W	Qiagen Blood Cell culture Minikit (modified)	16-Mar-18	Nanopore library	Dsilvatica_NMH1840_Nanopore	Nanopore 1D	-	long reads	-
NMH1829	*Dysdera silvatica*	Female	SAMN11609373	10-Mar-12	Teselinde, La Gomera, Spain	28.196 N	17.287 W	Qiagen Blood Cell culture Minikit (modified)	28-Jun-18	Nanopore library	Dsilvatica_NMH1829_Nanopore	Nanopore 1D	-	long reads	-

***Table S1-2*: DNA sequencing read files used in this study**

a) Short reads libraries

Library	Run ID	Read Lengths	Read #1 Filename	Read #2 Filename	Total Bases	Raw Read Pairs	GC (%)	Q20 (%)	Q30 (%)	NCBI SRA Accession
Short fragment	Dsilvatica_NMH2597_JFSH-CF_gDNA_PE	100x100 PE	Dsilvatica_NMH2597_JFSH-CF_gDNA_PE_1	Dsilvatica_NMH2597_JFSH-CF_gDNA_PE_2	51,202,445,102	506,954,902	36.68	95.26	86.55	SRR7340408
5 kb Nextera Mate Pair	Dsilvatica_NMH2605_JFSH-CF_gDNA_5kb_MP	100x100 PE	Dsilvatica_NMH2605_JFSH-CF_gDNA_5kb_MP_1	Dsilvatica_NMH2605_JFSH-CF_gDNA_5kb_MP_2	39,609,522,995	392,173,495	35.39	98.55	93.95	SRR7340407

b) PacBio long reads libraries

Library	Run ID	Read Lengths[a]	Read #1 Filename	Total Bases	Raw Read Pairs	NCBI SRA Accession
20 kb SMRTbell Templates	Dsilvatica_NMH2610_JFSH_PacB-1	SMRT	JFSH_NMH2610_Dsil_PB_subreads	1,268,701,492	175,819	SRR7429499
20 kb SMRTbell Templates	Dsilvatica_NMH2610_JFSH_PacB-2	SMRT	JFSH_NMH2610_Dsil_PB-2_subreads	1,262,187,088	185,427	SRR7429498
20 kb SMRTbell Templates	Dsilvatica_NMH2610_JFSH_PacB-3	SMRT	JFSH_NMH2610_Dsil_PB-3_subreads	1,215,789,613	184,206	SRR7429501
20 kb SMRTbell Templates	Dsilvatica_NMH2610_JFSH_PacB-4	SMRT	JFSH_NMH2610_Dsil_PB-4_subreads	1,185,146,691	179,94	SRR7429500
20 kb SMRTbell Templates	Dsilvatica_NMH2610_JFSH_PacB-5	SMRT	JFSH_NMH2610_Dsil_PB-5_subreads	1,209,437,458	186,657	SRR7429503
20 kb SMRTbell Templates	Dsilvatica_NMH2610_JFSH_PacB-6	SMRT	JFSH_NMH2610_Dsil_PB-6_subreads	1,231,887,854	194,049	SRR7429505
20 kb SMRTbell Templates	Dsilvatica_NMH2610_JFSH_PacB-7	SMRT	JFSH_NMH2610_Dsil_PB-7_subreads	1,240,408,666	193,592	SRR7429502
20 kb SMRTbell Templates	Dsilvatica_NMH2610_JFSH_PacB-8	SMRT	JFSH_NMH2610_Dsil_PB-8_subreads	1,039,286,018	155,598	SRR7429504
[a]SMRT: Single Molecule Real Time			Total PB subreads:	9,652,844,880	1,455,288	

c) Nanopore long read libraries

Library	Run ID	Read Lengths[b]	Read #1 Filename	Total Bases	Raw Read Pairs	NCBI SRA Accession
ONT_1	Dsilvatica_NMH1840_Nanopore_1	Nanopore	Dsilvatica_NMH1840_Nanopore_1.fastq	8.070.627.302	3517656	SRR9033390
ONT_2	Dsilvatica_NMH1840_Nanopore_2	Nanopore	Dsilvatica_NMH1840_Nanopore_2.fastq	5.672.731.340	2156287	SRR9033391
ONT_3	Dsilvatica_NMH1840_Nanopore_3	Nanopore	Dsilvatica_NMH1840_Nanopore_3.fastq	6.393.750.434	7407703	SRR9033392
ONT_4	Dsilvatica_NMH1829_Nanopore_4	Nanopore	Dsilvatica_NMH1829_Nanopore_4.fastq	1.808.992.949	1891482	SRR9033393
ONT_5	Dsilvatica_NMH1829_Nanopore_5	Nanopore	Dsilvatica_NMH1829_Nanopore_5.fastq	1.247.255.456	1312358	SRR9033389
[a] Nanopore variable lengths			Total Nanopore subreads:	23,193,357,481	16,285,486	

Table S1-3: **NCBI Data used for the contaminant search step**

Taxon	Genomes retrieved	FTP site
Archae	683	ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/archaea/assembly_summary.txt
Virus	7.538	ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/viral/assembly_summary.txt
Bacteria	95.973	ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/assembly_summary.txt
Mitochondria	5.883	ftp://ftp.ncbi.nlm.nih.gov/refseq/release/mitochondrion/

Downloaded on October, 2015

Table S1-4: **Pre-processing statistics for each library**

a) Trimming of PE reads using Trimmomatic

Class	Nº reads	Percentage (%)
Input read pairs	253.445.279	100
Both surviving	243.763.790	96,18
Forward only	8.031.776	3,17
Reverse only	1.312.067	0,52
Dropped	337.646	0,13

b) Trimming of MP reads using NxTrim

Nº reads	Percentage (%)	
195.996.473	100	reads passed chastity/purity filters
189	0,00	reads had two copies of adapter
166.336	0,08	reads pairs were ignored because a template length appeared less than read length
195.829.948	99,91	remaining reads were trimmed
62.192.088	31,76	read pairs had MP orientation
50.195.582	25,63	read pairs had PE orientation
81.678.049	41,71	read pairs had Unknown orientation
1.764.229	0,90	were single end reads
23.410.880	11,95	extra single end reads were generated from overhangs

c) Circularizing Pacbio reads using SMRT analysis software

Lane	Raw Bases	Subreads	Circularized Bases	CCS reads*
1	1.268.701.492	175.819	46.040.000	5.184
2	1.262.187.088	185.427	62.310.000	7.364
3	1.215.789.613	184.206	65.070.000	7.966
4	1.185.146.691	179.940	64.960.000	7.796
5	1.209.437.458	186.657	73.500.000	8.772
6	1.231.887.854	194.049	58.240.000	7.954
7	1.240.408.666	193.592	52.170.000	7.437
8	1.039.286.018	155.598	33.740.000	4.824
TOTAL	9.652.844.880	1.455.288	456.030.000	57.297

CCS: Circularized Consensus Sequences

Table S1-5: **Samples used for the RNAseq study**

Species	Sex	Sampling date	Latitude	Longitude	Location	Tissue	Library	Machine	NCBI BioSample Accession
Dysdera silvatica	Male	2013	28.112 N	17.262 W	Las Tajoras, La Gomera, Spain	First pair of legs	RNA TrueSeq	Illumina HiSeq 2000	SAMN04527047
Dysdera silvatica	Male	2013	28.112 N	17.262 W	Las Tajoras, La Gomera, Spain	Pedipalps	RNA TrueSeq	Illumina HiSeq 2000	SAMN04527048
Dysdera silvatica	Male	2013	28.112 N	17.262 W	Las Tajoras, La Gomera, Spain	Remaining legs	RNA TrueSeq	Illumina HiSeq 2000	SAMN04527049
Dysdera silvatica	Male	2013	28.112 N	17.262 W	Las Tajoras, La Gomera, Spain	Remaining parts of the spider	RNA TrueSeq	Illumina HiSeq 2000	SAMN04527050

All samples belong to the NCBI project PRJNA313901 and they are all a mixed of the RNA extraction from four different *D. silvatica* individuals with code: NMH2597, NMH2598, NMH2599, NMH2601

Reference:

Joel Vizueta, Cristina Frías-López, Nuria Macías-Hernández, Miquel A. Arnedo, Alejandro Sánchez-Gracia, Julio Rozas;
Evolution of Chemosensory Gene Families in Arthropods: Insight from the First Inclusive Comparative Transcriptome Analysis across Spider Appendages.
Genome Biol Evol 2017; 9 (1): 178-196. doi: 10.1093/gbe/evw296

***Table S1-6*: RNA sequencing read files**

Library	Read Lengths	Tissue	Read #1 Filename	Read #2 Filename	# Read Pairs	NCBI SRA Accession
RNA TrueSeq	100x100 PE	First pair of legs	CF1Leg_1	CF1Leg_2	59.008.693	SRX1612801
RNA TrueSeq	100x100 PE	Pedipalps	CF2Palp_1	CF2Palp_2	57.493.091	SRX1612802
RNA TrueSeq	100x100 PE	Remaining legs	CF3Rest_1	CF3Rest_1	51.932.520	SRX1612803
RNA TrueSeq	100x100 PE	Remaining parts of the spider	CF4Body_1	CF4Body_2	52.483.628	SRX1612804

All samples belong to the NCBI project PRJNA313901 and they are all a mixed of the RNA extraction from four different *D. silvatica* individuals with code: NMH2597, NMH2598, NMH2599, NMH2601

Reference:

Joel Vizueta, Cristina Frías-López, Nuria Macías-Hernández, Miquel A. Arnedo, Alejandro Sánchez-Gracia, Julio Rozas;
Evolution of Chemosensory Gene Families in Arthropods: Insight from the First Inclusive Comparative Transcriptome Analysis across Spider Appendages.
Genome Biol Evol 2017; 9 (1): 178-196. doi: 10.1093/gbe/evw296

***Table S1-7*: Genome descriptive statistics**

a) Summary of the comparison of genome assembly steps

		MaSurCA Assembly	AGOUTI RNA scaffolding	Clean contaminants (v.1.2)
Assembly Statistics	Sequences	66.830	65.221	65.205
	Total Length (bp)	1.359.506.718	1.361.115.718	1.359.336.805
	A	32,484	32,445	32,445
	T	32,504	32,468	32,467
	C	17,431	17,412	17,411
	G	17,44	17,417	17,417
	A+T	64,989	64,912	64,913
	C+G	34,87	34,829	34,828
	N	0,14	0,258	0,258
	Length (no Ns)	1.357.599.280	1.357.599.280	1.355.820.767
	Capture Gaps[a]	19.075	20.684	20.680
	Capture Gaps Length	1.907.438	3.516.438	3.516.038
	MinLen	1.001	1.001	1.001
	MaxLen	381.119	381.119	340.047
	Average Len	20.342,76	20.869,29	20.847,13
	Median Len	12.186	12.176	12.171
	n50	36.307	38.050	38.017
	L50	11.019	10.428	10.436
Completeness statistics	BUSCO Arthropoda[b] [*n* = 1066]	Identified (I)	879 (82.46%)	916 (85.93%)	921 (86.4%)
		Complete (C)	739 (69.3%)	786 (73.7%)	782 (73.4%)
		Complete and single copy (S)	709 (66.5%)	753 (70.6%)	750 (70.4)
		Complete and duplicated (D)	30 (2.8%)	33 (3.1%)	32 (3%)
		Fragmented (F)	140 (13.1%)	130 (12.2%)	139 (13%)
		Missing (M)	187 (17.6%)	150 (14.1%)	145 (13.6%)
	BUSCO Metazoa[b] [*n* = 978]	Identified (I)	818 (83.64%)	855 (87.42%)	855 (87.42%)
		Complete (C)	689 (70.5%)	739 (75.6%)	734 (75.1%)
		Complete and single copy (S)	661 (67.6%)	705 (72.1%)	706 (72.2%)
		Complete and duplicated (D)	28 (2.9%)	34 (3.5%)	28 (2.9%)
		Fragmented (F)	129 (13.2%)	116 (11.9%)	121 (12.4%)
		Missing (M)	160 (16.3%)	123 (12.5%)	123 (12.5%)
	BLAST[f]	1 to 1 Dysdera[c] [9473]	9,121 (96.28%)	9,121 (96.28%)	9,120 (96.27%)
		CEGMA[d] [457]	435 (95.19%)	435 (95.19%)	434 (94.97%)
		Dysdera silvatica Transcripts[e] [58966]	40,778 (69.16%)	40,771 (69.14%)	40,499 (68.68%)

[a] Capture gaps are defined as regions containing a minimum of 25 consecutive base pairs codified as "N".

[b] BUSCO analysis using default parameters for different datasets: arthropoda (*n* = 1066 genes) and metazoa (*n* = 978 genes)

[c] TBLASTX search statistics (*evalue* 1e-30) of 1:1 Ortholog proteins (9,473) identified in the species *D. silvatica*, *D. gomerensis* Strand, 1911; *D. verneaui* Simon, 1883; *D. tilosensis* Wunderlich, 1992 and *D. bandamae* Schmidt, 1973 (Joel Vizueta et al., 2019, unpublished results)

[d] TBLASTX search statistics (evalue 1e-5) of CEGMA proteins from *Drosophila melanogaster* (*n* = 457).

[e] TBLASTX search statistics (evalue: 1e-30) of proteins from *Dysdera silvatica* transcriptome (*n* = 58,966) (Vizueta et. al 2017).

[f] Identified hits within the genome at the give e-value for each dataset at any coverage

b) Genome annotation statistics[a]

Protein coding genes	48.619
Functionally annotated	36,398 (74.86%)
Swissprot	17,225 (35.43%)
InterPro[b]	32,322 (66.48%)
Without functional annotation	12,221 (25.14%)
tRNA genes	33.934

[a] Values are refered to final assembly version (v1.2)

[b] Analysis using the following databases (included in InterProScan): PANTHER (Mi, Muruganujan, and Thomas 2013), Pfam (Punta et al. 2012), PRINTS (Attwood et al. 2012), PrositePatterns & PrositeProfiles (Gattiker, Gasteiger, and Bairoch 2002; Sigrist et al. 2012), SMART (Letunic, Doerks, and Bork 2012), SUPERFAMILY (de Lima Morais et al. 2011), TIGRAM (Haft et al. 2012), SFLD (Akiva et al. 2014), Gene3D (Lees et al. 2012), Hamap (Pedruzzi et al. 2015), ProDom (Bru et al. 2004), PIRSF (Nikolskaya et al. 2007), MobiDBLite (Necci et al. 2017).

Table S1-8: Coverage analyses

a) Estimated coverage based on the total sequencing data

Run ID	Total Bases	Coverage[a]
PE	51.202.445.102	30X
MP	39.609.522.995	23X
ONT	23.193.357.481	14X
Pacbio	9.652.844.880	6X
Pacbio CSS	456.030.000	0.27X
All	123.658.170.458	73X

[a]Based on the genome size estimated by flow cytometry (1.7 Gb)
ONT: Oxford Nanoporte Technologies

b) Estimated coverage based on mapped data

Run ID	Mean Coverage[a]	Mean Coverage[b]
PE	25.7X	32.12X
MP	12.58X	15.72X
ONT	8.46X	10.57X
Pacbio CSS	0.23X	0.23X
All	46.97X	58.64X

[a]Based on the genome size estimated by flow cytometry (1.7 Gb)
[b]Based on the genome size assembled (1.36 Gb)
ONT: Oxford Nanoporte Technologies

Table S1-9: **RepeatMasker analysis of *D. silvatica* genome**

a) RepeatMasker analyses

Type	Number of elements[a]	Length (bp)	Sequence (%)
SINEs[b]	170.863	25.173.255	1,85
LINEs[c]	256.091	145.456.128	10,7
LINE1	608	152.963	0,01
LINE2	37.773	21.216.392	1,56
L3/CR1	2.525	1.461.178	0,11
LTR elements[d]	20.021	8.746.088	0,64
ERV_classI	171	155.175	0,01
ERV_classII	7.604	940.768	0,07
DNA elements	779.436	227.842.465	16,76
hAT-Charlie	119.132	24.177.681	1,78
TcMar-Tigger	67.086	12.923.815	0,95
Unclassified:	1.729.161	306.272.536	22,53
Total interspersed repeats		713.490.472	52,48
Small RNA	665	960.355	0,07
Satellites	4.653	1.491.285	0,11
Simple repeats	282.910	13.440.971	0,99
Low complexity	41.169	2.157.298	0,16
Total Bases masked (bp)	**3.284.969**	**731.540.381**	**53,81**

[a]Most repeats fragmented by insertions or deletions have been counted as one element
[b] SINEs: short interspersed nuclear element
[c] LINEs: long interspersed nuclear element
[d] LTR: long terminal repeat

b) Top 10 Repeat families analysis[a]

Family name	Count hits[b]	Type of family	Count total hits (%)[c]	Mean length[d]	Standard Deviation	Mean length Unknown[d]	Mean length SINEs[d]	Mean length LINEs[d]
rnd-2_family-13	28.831	Unknown	0,915	179,53	108,72	179,53	-	-
rnd-6_family-515	25.708	Unknown	0,816	257,79	203,56	257,79	-	-
rnd-1_family-23	23.184	Unknown	0,736	122,18	40,65	122,18	-	-
rnd-3_family-426	22.537	Unknown	0,715	158,04	88,64	158,04	-	-
rnd-1_family-35	20.382	LINE/RTE-RTE	0,647	1412,43	1047,77	-	-	1412,43
rnd-4_family-547	20.258	SINE/ID	0,643	119,51	64,39	-	119,51	-
rnd-1_family-36	18.802	Unknown	0,597	251,98	325,65	251,98	-	-
rnd-1_family-7	18.350	SINE/ID	0,582	212,24	49,16	-	212,24	-
rnd-3_family-155	17.829	LINE/RTE-RTE	0,566	665,48	654,55	-	-	665,48
rnd-6_family-1201	16.467	SINE/ID	0,523	152,32	97,43	-	152,32	-
Total	212.348		6.741			193,91	161,36	1038,96

[a] Top 10 most common repeats across the 2,604 families identified
[b]The number of times (hits) identified across the genome
[c] Percentage of hits for each family identified among the 3,150,262 total hits identified
[d] Mean repeat length

***Table S1-10*: Reference guided transcriptome assembly statistics**

Sequences	245.905
Total Length (bp)	145.213.967
A (%)	31,867
T (%)	31,599
C (%)	18,049
G (%)	18,483
A+T (%)	63,467
C+G (%)	36,532

Set	>150 bp	>500 bp	>1000 bp
Number Seqs	245.905	79.722	33.778
% Seqs	100	32,419	13,736
Total Length (bp)	145.213.967	94.768.228	63.221.799
% Bases	100	65,261	43,537
Minimum Length (bp)	201	501	1.001
Maximun Length (bp)	16.350	16.350	16.350
Average Length (bp)	590,53	1.188,73	1.871,69
Median Length (bp)	355	875	1.547
N50	804	1.392	1.956

Table S1-11: Source of proteins to conduct the annotation of _D. silvatica_ genes and completeness analysis (Figure 2).

Species	Taxonomic range	Source	Downloaded from (id)
Drosophila melanogaster	hexapoda	UNIPROT	ftp uniprot.org (UP000000803)
Bombyx mori	hexapoda	UNIPROT	ftp uniprot.org (UP000005204)
Pediculus humanus	hexapoda	UNIPROT	ftp uniprot.org (UP000009046)
Daphnia pulex	crustacea	UNIPROT	ftp uniprot.org (UP000000305)
Strigamia maritima	myriapoda	UNIPROT	ftp uniprot.org (UP000014500)
Hypsibius dujardini	tardigrada	UNIPROT	ftp uniprot.org (UP000192578)
Caenorhabditis elegans	nematoda	UNIPROT	ftp uniprot.org (UP000001940)
Tetranychus urticae	acari	UNIPROT	ftp uniprot.org (UP000015104)
Euroglyphus maynei	acari	UNIPROT	ftp uniprot.org (UP000194236)
Ixodes scapularis	acari	UNIPROT	ftp uniprot.org (UP000001555)
Metaseiulus occidentalis	acari	NCBI	ftp ncbi.nlm.nih.gov (GCF_000255335.1_Mocc_1.0)
Stegodyphus mimosarum	aranae	UNIPROT	ftp uniprot.org (UP000054359)
Acanthoscurria geniculata	aranae	Original paper[a]	Additional information original paper
Lactodrectus hesperus	aranae	i5K	i5K Project
Loxosceles reclusa	aranae	i5K	i5K Project
Parasteatoda tepidariorum	aranae	i5K	i5K Project
Mesobuthus martensii	scorpion	NCBI	ftp ncbi.nlm.nih.gov (GCA_000484575.1)
Centruroides sculpturatus	scorpion	i5K	i5K Project
Limulus polyphemus	Xiphosura	RyanLab website	ryanlab.whitney.ufl.edu (GCF_000517525.1_Limulus_polyphemus-2.1.2)

[a] https://doi.org/10.1038/ncomms4765

***Table S1-12*: Annotation statistics**

a) Cumulative frecuence distribution of the Annotation Edit Distance (AED) for each set

		Annotation Round					
		R1	R2	F1	F2	Functional Annotation	Unique Dsilvatica
	Mean	0,374	0,36	0,323	0,32	0,268	0,401
Cumulative Frequence							
[0	0.02)	0,0335	0,04	0,0642	0,048	0,0642	0,0405
[0.02	***0.04)***	0,0439	0,0528	0,0921	0,0765	0,0921	0,0584
[0.04	0.06)	0,0565	0,0691	0,1199	0,1043	0,1199	0,077
[0.06	***0.08)***	0,0713	0,0889	0,1468	0,132	0,1468	0,0947
[0.08	0.1)	0,0892	0,1115	0,174	0,1607	0,174	0,1099
[0.1	***0.12)***	0,1081	0,1358	0,2014	0,1894	0,2014	0,1261
[0.12	0.14)	0,1277	0,1609	0,2303	0,2189	0,2303	0,1406
[0.14	***0.16)***	0,1507	0,1897	0,2591	0,2484	0,2591	0,1588
[0.16	0.18)	0,1745	0,2214	0,2871	0,278	0,2871	0,1753
[0.18	***0.2)***	0,1987	0,2521	0,3174	0,3098	0,3174	0,1951
[0.2	0.22)	0,2265	0,2848	0,3475	0,3407	0,3475	0,217
[0.22	***0.24)***	0,2545	0,3162	0,3787	0,3739	0,3787	0,242
[0.24	0.26)	0,2874	0,3526	0,4089	0,4053	0,4089	0,2693
[0.26	***0.28)***	0,3219	0,3898	0,4404	0,4386	0,4404	0,296
[0.28	0.3)	0,3676	0,4301	0,4748	0,4731	0,4748	0,3321
[0.3	***0.32)***	0,4044	0,4675	0,51	0,5079	0,51	0,3609
[0.32	0.34)	0,4424	0,505	0,5466	0,5459	0,5466	0,3943
[0.34	***0.36)***	0,4823	0,543	0,5834	0,583	0,5834	0,4269
[0.36	0.38)	0,5223	0,5797	0,6197	0,6211	0,6197	0,463
[0.38	***0.4)***	0,5656	0,6177	0,6567	0,6594	0,6567	0,5018
[0.4	0.42)	0,6074	0,6535	0,6945	0,7	0,6945	0,5497
[0.42	***0.44)***	0,6493	0,6868	0,7302	0,7372	0,7302	0,5921
[0.44	0.46)	0,6917	0,7183	0,7637	0,7738	0,7637	0,6287
[0.46	***0.48)***	0,7345	0,7471	0,7927	0,8063	0,7927	0,6641
[0.48	0.5)	0,7748	0,7713	0,8157	0,8331	0,8157	0,6985
[0.5	***0.52)***	0,7936	0,7863	0,8312	0,8492	0,8312	0,7225
[0.52	0.54)	0,81	0,8004	0,8455	0,8629	0,8455	0,7451
[0.54	***0.56)***	0,825	0,8135	0,8583	0,8754	0,8583	0,7625
[0.56	0.58)	0,84	0,8267	0,8694	0,8866	0,8694	0,7782
[0.58	***0.6)***	0,854	0,8392	0,8805	0,8971	0,8805	0,7991
[0.6	0.62)	0,8681	0,851	0,891	0,9061	0,891	0,8153
[0.62	***0.64)***	0,8806	0,8623	0,9013	0,9157	0,9013	0,8337
[0.64	0.66)	0,8927	0,8732	0,9106	0,924	0,9106	0,8504
[0.66	***0.68)***	0,9047	0,884	0,9195	0,932	0,9195	0,8644
[0.68	0.7)	0,9204	0,9005	0,9317	0,943	0,9317	0,8865
[0.7	***0.72)***	0,9256	0,9051	0,9353	0,9462	0,9353	0,8919
[0.72	0.74)	0,9358	0,9155	0,9429	0,9533	0,9429	0,9054
[0.74	***0.76)***	0,9453	0,9262	0,9503	0,9591	0,9503	0,9165
[0.76	0.78)	0,95389	0,9351	0,9571	0,9647	0,9571	0,9273
[0.78	***0.8)***	0,9617	0,9443	0,963	0,9696	0,963	0,9378
[0.8	0.82)	0,9713	0,9574	0,971	0,9764	0,971	0,9491
[0.82	***0.84)***	0,9742	0,9614	0,9732	0,9782	0,9732	0,9543
[0.84	0.86)	0,9797	0,9684	0,9785	0,9822	0,9785	0,9643
[0.86	***0.88)***	0,9844	0,9753	0,9827	0,9855	0,9827	0,9705
[0.88	0.9)	0,9884	0,982	0,9867	0,9889	0,9867	0,9783
[0.9	***0.92)***	0,9919	0,9869	0,9901	0,9916	0,9901	0,9837
[0.92	0.94)	0,9957	0,9933	0,9948	0,9953	0,9948	0,9911
[0.94	***0.96)***	0,9968	0,9949	0,996	0,9965	0,996	0,9923
[0.96	0.98)	0,9986	0,9977	0,9981	0,9984	0,9981	0,9955
[0.98	***1)***	1	1	1	1	1	1

b) Annotation metrics

	Number	Average Exons	Average Length (bp)	Average 5' UTR (bp)	Average 3' UTR (bp)
All proteins	48.619	3,82	221,357	27,7	79,32
Functional annotated	36.398	4,11	258,294	31,41	97,48
Unique Dsilvatica	4.077	2,81	168,61	13,89	26,49

c) Values of the OrthoDB groups obtained

Chelicerata (Figure 4b)

Species	Single copy in All species	Single copy in All-but-1	Copies in All	Copies in All-but-1	Other orthology
Dysdera silvatica	1.798	952	2.212	656	8.281
Parasteatoda tepidariorum	1.798	1.191	2.212	782	9.584
Stegodyphus mimosarum	1.798	1.144	2.212	753	9.713
Ixodes scapularis	1.798	988	2.212	614	7.823
Tetranychus urticate	1.798	649	2.212	443	5.994

Arthropods (Figure S8)

Species	Single copy in All species	Single copy in All-but-1	Copies in All	Copies in All-but-1	Other orthology
Dysdera silvatica	950	588	2.623	1.091	7.116
Parasteatoda tepidariorum	950	691	2.623	1.208	8.977
Ixodes scapularis	950	598	2.623	1.040	8.223
Limulus polyphemus	950	690	2.623	1.217	9.712
Drosophila melanogaster	950	435	2.623	640	9.642
Strigamia maritima	950	568	2.623	1.004	7.491

Table S1-13: Mitochondrial assembly metrics and annotation features

a) Assembly metrics

Sequences	1
Length	14,440 bp
Reads PE library	126.758
Average Coverage	878X

b) Annotation

Reference	Software	Type	Name	Start Coordinates	End Coordinates	Strand
Contig1	mitfi	tRNA	trnI(atc)	466	534	+
Contig1	mitos	gene	nad2	552	1.349	+
Contig1	mitfi	tRNA	trnW(tga)	1.450	1.515	+
Contig1	mitfi	tRNA	trnC(tgc)	1.527	1.573	-
Contig1	mitos	gene	cox1	1.569	3.080	+
Contig1	mitos	gene	cox2	3.163	3.762	+
Contig1	mitfi	tRNA	trnD(gac)	3.849	3.902	+
Contig1	mitos	gene	atp8	3.896	4.033	+
Contig1	mitos	gene	atp6	4.045	4.695	+
Contig1	mitos	gene	cox3	4.735	5.490	+
Contig1	mitfi	tRNA	trnG(gga)	5.492	5.556	+
Contig1	mitos	gene	nad3	5.538	5.867	+
Contig1	mitfi	tRNA	trnL2(tta)	5.858	5.913	-
Contig1	mitfi	tRNA	trnN(aac)	5.995	6.059	+
Contig1	mitfi	tRNA	trnS1(agc)	6.058	6.117	+
Contig1	mitfi	tRNA	trnF(ttc)	6.181	6.243	-
Contig1	mitos	gene	nad5	6.263	7.840	-
Contig1	mitfi	tRNA	trnH(cac)	7.887	7.935	-
Contig1	mitos	gene	nad4	7.940	9.211	-
Contig1	mitos	gene	nad4l	9.216	9.461	-
Contig1	mitfi	tRNA	trnP(cca)	9.461	9.518	-
Contig1	mitos	gene	nad6	9.522	9.944	+
Contig1	mitos	gene	cob	9.964	11.076	+
Contig1	mitfi	tRNA	trnS2(tca)	11.078	11.130	+
Contig1	mitfi	tRNA	trnT(aca)	11.135	11.183	+
Contig1	mitos	gene	nad1	11.203	12.081	-
Contig1	mitfi	tRNA	trnL1(cta)	12.076	12.143	-
Contig1	mitfi	rRNA	rrnL	12.156	12.714	-
Contig1	mitfi	tRNA	trnV(gta)	13.118	13.174	-
Contig1	mitfi	rRNA	rrnS	13.174	13.933	-

***Table S2-1*: Example of results for the HCR analysis**

a) Example data file including the high coverage regions (HCRs) data[a]

Scaffold ID	Min Intra length[b]	Scaffold Length	# HCR[c]	HCR Id[d]	Inter Start[e]	Intert End[e]	Inter Length[f]	Intra Start[g]	Intra End[g]	Intra Length[h]
sequence_10003	150	24752	4	repeat_1	-	-	-	6338	6508	170
sequence_10003	150	24752	4	repeat_2	6509	11083	4574	11084	11776	692
sequence_10003	150	24752	4	repeat_3	11777	20259	8482	20260	21098	838
sequence_10003	150	24752	4	repeat_4	21099	22039	940	22040	22497	457
sequence_10005	150	13790	3	repeat_1	-	-	-	9668	10197	529
sequence_10005	150	13790	3	repeat_2	10198	10335	137	10336	11212	876
sequence_10005	150	13790	3	repeat_3	11213	11250	37	11251	11749	498
sequence_10011	150	14044	1	repeat_1	-	-	-	8016	8201	185
sequence_10010	150	12713	2	repeat_1	-	-	-	449	1347	898
sequence_10010	150	12713	2	repeat_2	1348	1484	136	1485	1715	230
sequence_10008	150	78454	2	repeat_1	-	-	-	56694	56874	180
sequence_10008	150	78454	2	repeat_2	56875	74155	17280	74156	74466	310
...	...	...	...	...	...	...	...	...	...	...

[a] Example of the first entries of the data file (subset_10kb-2.5x_coverage.HCR_150.txt). This data file contains full information and is provided as Supplementary Files.

[b] Minimum intra-repeat length cutoff

[c] Number of High Coverage regions (HCR) in this scaffold

[d] HCR identifier within scaffold

[e] Inter-repeat start and end coordinates within the scaffold

[f] Inter-repeat length

[g] Intra-repeat start and end coordinates within the scaffold

[h] Intra-repeat length

b) Example in BED format of the intersection of the annotation between RepeatMasker and HCR data[a]

Scaffold ID	Intra HCR Start	Intra HCR End	HCR ID	HCR Score[b]	HCR Strand[c]	Scaffold ID	Repeat Start	Repeat End	Repeat (Family == Type)	Repeat Score	Repeat Strand	Overlapping Base Pairs[d]
sequence_10003	11084	11776	repeat_2	1	.	sequence_10003	10627	11807	rnd-4_family-152==LINE/RTE-RTE	10056	+	692
sequence_10003	20260	21098	repeat_3	1	.	sequence_10003	19683	22063	rnd-1_family-685==LINE/L2	13611	+	838
sequence_10003	22040	22497	repeat_4	1	.	sequence_10003	19683	22063	rnd-1_family-685==LINE/L2	13611	+	23
sequence_10005	10336	11212	repeat_2	1	.	sequence_10005	9655	10691	rnd-1_family-678==Unknown	8946	+	355
sequence_10005	10336	11212	repeat_2	1	.	sequence_10005	10283	11237	rnd-1_family-10==Unknown	8156	+	876
sequence_10005	10336	11212	repeat_2	1	.	sequence_10005	10828	11768	rnd-1_family-364==Unknown	7436	+	384
sequence_10005	11251	11749	repeat_3	1	.	sequence_10005	10828	11768	rnd-1_family-364==Unknown	7436	+	498
sequence_10010	1485	1715	repeat_2	1	.	sequence_10010	1	1804	rnd-1_family-84==LINE/Dong-R4	15595	+	230
sequence_10008	74156	74466	repeat_2	1	.	sequence_10008	74123	74520	rnd-1_family-84==LINE/Dong-R4	3457	+	310
...	...	...	...	...	...	...	...	...	...	...	...	...

[a] Example of the firsts entries of the data file (HCR_annotation_2.5_150-intersection.bed). This data file was generated using Bedtools (intersectBed -wao option) to calculate the intersection of the annotation by repeatmasker and the HCR annotation previously generated. The whole data file is provided as Supplementary Files

[b] HCR Score: BED format attribute. Check for further details https://genome.ucsc.edu/FAQ/FAQformat.html#format1

[c] HCR Strand: No information available.

[d] Overlapping base pairs among both features (Repeatmasker annotation and HCR element)

***Table S2-2*: High coverage regions (HCRs) analysis**

a) Descriptive statistics in the subset (34,937 contigs)

	Length	Number HCR[a]	Contigs containing HCR (%)	HCR/Contig[b]	HCR/Total Contig[c]	Number HCR/Total Mb
2.5x	150	86.641	21,614 (61.87)	4,01	2,48	77,7
	500	18.375	7,764 (22.22)	2,37	0,53	16,48
	1.000	5.637	2,605 (7.45)	2,16	0,16	5
	5.000	251	183 (0.52)	1,37	0,01	0,23
5x	150	28.512	10,604 (30.35)	2,69	0,81	25,57
	500	4,690	2,277 (6.51)	2,06	0,12	4,21
	1.000	1,248	628 (1.79)	1,98	0,03	1,12
	5.000	31	26 (0.07)	1,19	0	0,03

[a] High Cover

[b] Median number of HCR per contig containing ≥1 HCR

[c] Median number of HCR per contig.

b) Inter and Intra HCR lengths distribution

Intra-HCR

	Length	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
2.5x	150	150	195	274	433,6	450	21.901
	500	500	597	755	1.079	1.120	21.901
	1.000	1.000	1.170	1.467	1.985	2.195	21.901
	5.000	5.016	5.447	6.243	7.119	8.200	21.901
5x	150	150	192	261	377,7	398	11.598
	500	500	580	712,5	980,9	1029,8	11.598
	1.000	1.000	1.162	1.436	1.850	2.048	11.598
	5.000	5.013	5.770	7.170	7.060	7.826	11.598

Inter-HCR

	Length	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
2.5x	150	10	149	523	4.081	2.691	144.900
	500	10	161	666	3.817	2.483	150.879
	1.000	10	163	855,5	2.782,30	2.577,50	82.821
	5.000	20	272	3.375	7.389	12.348	31.256
5x	150	10	161	504	3.901	1.896	146.087
	500	10	189,2	768,5	2.882,20	2.462,80	81.329
	1.000	10	197,8	1.037	2.762,30	3.035	41.244
	5.000	232	476,5	892,5	1.938,50	2.354,50	5.737

***Table S2-3*: Enrichment analysis of the intersection of High Coverage regions (HCRs) with RepeatMasker annotation (main repeats)[a]**

a) 2.5x Mean Coverage

Type	Average Length (bp)	Genome[b]	> 150 bp	> 500 bp	> 1000 bp	> 5000 bp
SINEs[d]	157,13	187240 (5.87 %)[a]	3552 (4.07 %) [0/1][c]	1820 (4.93 %) [0/1]	1238 (5.75 %) [0.389/0.622]	218 (6.14 %) [0.769/0.253]
LINEs[e]	541,12	274771 (8.62 %)	31478 (36.12 %) [1/0]	7651 (20.74 %) [1/0]	2601 (12.08 %) [1/0]	355 (10.01 %) [0.998/0.002]
LTR elements[f]	388,72	22969 (0.72 %)	1288 (1.48 %) [1/0]	493 (1.34 %) [1/0]	238 (1.11 %) [1/0]	34 (0.96 %) [0.957/0.062]
DNA elements	283,14	824501 (25.86 %)	18126 (20.8 %) [0/1]	8466 (22.94 %) [0/1]	5297 (24.61 %) [0/1]	881 (24.84 %) [0.085/0.921]
Unclassified	176,24	1840236 (57.72 %)	31928 (36.63 %) [0/1]	18022 (48.84 %) [0/1]	11847 (55.04 %) [0/1]	2004 (56.5 %) [0.073/0.931]
Small RNA[g]	1.128,62	851 (0.03 %)	261 (0.3 %) [1/0]	92 (0.25 %) [1/0]	56 (0.26 %) [1/0]	0 (0 %) [0.388/1]
Satellites	307,13	4890 (0.15 %)	166 (0.19 %) [0.9974/0.0034]	76 (0.21 %) [0.995/0.008]	39 (0.18 %) [0.87/0.168]	6 (0.17 %) [0.696/0.461]
Low complexity	51,2	41405 (1.3 %)	422 (0.48 %) [0/1]	314 (0.85 %) [0/1]	229 (1.06 %) [0.001/0.999]	54 (1.52 %) [0.893/0.136]
	Total	3,188,449[h]	87.159	36.898	21.526	3.547

b) 5x Mean Coverage

Type	Average Length (bp)	Genome[b]	> 150 bp	> 500 bp	> 1000 bp	> 5000 bp
SINEs[d]	157,13	187240 (5.87 %)	948 (4.29 %) [0/1]	471 (5.72 %) [0.282/0.734]	251 (6.11 %) [0.756/0.265]	21 (5.36 %) [0.383/0.698]
LINEs[e]	541,12	274771 (8.62 %)	8305 (37.6 %) [1/0]	1434 (17.4 %) [1/0]	466 (11.35 %) [1/0]	42 (10.71 %) [0.938/0.085]
LTR elements[f]	388,72	22969 (0.72 %)	255 (1.15 %) [1/0]	72 (0.87 %) [0.953/0.06]	35 (0.85 %) [0.862/0.18]	0 (0 %) [0.059/1]
DNA elements	283,14	824501 (25.86 %)	3939 (17.83 %) [0/1]	1835 (22.27 %) [0/1]	1020 (24.84 %) [0.07/0.935]	105 (26.79 %) [0.686/0.356]
Unclassified	176,24	1840236 (57.72 %)	8191 (37.08 %) [0/1]	4212 (51.12 %) [0/1]	2226 (54.21 %) [0/1]	216 (55.1 %) [0.16/0.864]
Small RNA[g]	1.128,62	851 (0.03 %)	243 (1.1 %) [1/0]	95 (1.15 %) [1/0]	44 (1.07 %) [1/0]	0 (0 %) [0.901/1]
Satellites	307,13	4890 (0.15 %)	107 (0.48 %) [1/0]	47 (0.57 %) [1/0]	19 (0.46 %) [1/0]	1 (0.26 %) [0.878/0.452]
Low complexity	51,2	41405 (1.3 %)	123 (0.56 %) [0/1]	81 (0.98 %) [0.005/0.996]	47 (1.14 %) [0.213/0.826]	7 (1.79 %) [0.858/0.25]
	Total	3,188,449[h]	22.089	8.240	4.106	392

[a]Results generated using Bedtools (analysis of the intersection of the annotation by repeatmasker and the HCR annotations). The number of hits for each element could slightly differ from those results reported in Table S1-9a, as in this analysis we have not take into account overlapping annotations, or multiple annotations in the same region.

[b]Count -number of elements- and percentage across repeat elements (%)

[c]Count (%) [CDF/SF]; where CDF and SF are the P-values

CDF = *P*-value of observing (obtaining by random) a number of repeat elements less than or equal to the observed

SF = *P*-value of observing (obtaining by random) a number of repeat elements greater than or equal to the observed

[d]SINEs: short interspersed nuclear element

[e]LINEs: long interspersed nuclear element

[f]LTR: long terminal repeat

[g]Small RNA referes to rRNA

[h]Total amount of identified repeats could slightly differ from those reported by RepeatMasker. In the current analysis we have not take into account overlapping annotations, or multiple annotations in the same region.

Table S2-4: Enrichment analysis of the intersection of High Coverage regions (HCRs) (2.5 x mean coverage) with RepeatMasker annotation (subtypes of main repeats)[a]

				> 150 bp					> 500 bp					> 1000 bp					> 5000 bp				
	Total Genome[b]	Total Genome type[c]	%	Total Count Subset[d]	Total Subset Type[e]	%	CDF[f]	SF[g]	Total Count Subset	Total Subset Type	%	CDF	SF	Total Count Subset	Total Subset Type	%	CDF	SF	Total Count Subset	Total Subset Type	%	CDF	SF
DNA/CMC-Chapaev-3	3188449	5416	0,170	87159	127	0,146	0,041	0,966	36898	67	0,182	0,734	0,308	21526	38	0,177	0,635	0,428	3547	10	0,282	0,957	0,085
DNA/CMC-EnSpm	3188449	10378	0,32549	87159	198	0,22717	3,08514E-08	0,999999975	36898	100	0,27102	0,033011047	0,973556235	21526	48	0,22299	0,003260787	0,99785522	3547	7	0,19735	0,1108	0,9416
DNA/CMC-Transib	3188449	1359	0,043	87159	40	0,046	0,718	0,340	36898	17	0,046	0,685	0,407	21526	12	0,056	0,863	0,214	3547	1	0,028	0,554	0,780
DNA/Crypton-H	3188449	1430	0,04485	87159	27	0,03098	0,02519494	0,98378469	36898	15	0,04065	0,412505741	0,682982943	21526	12	0,05575	0,823886197	0,264128302	3547	2	0,05639	0,7858	0,4722
DNA/Ginger	3188449	9241	0,290	87159	159	0,182	0,000	1,000	36898	81	0,220	0,005	0,996	21526	40	0,186	0,002	0,999	3547	11	0,310	0,665	0,452
DNA/hAT	3188449	364071	11,4184	87159	5891	6,75891	0	0,999999992	36898	3216	8,71592	2,40773E-65	0,999999998	21526	2113	9,81604	1,96349E-14	0,999999995	3547	380	10,7133	0,0969	0,912
DNA/hAT-Ac	3188449	7724	0,242	87159	153	0,176	0,000	1,000	36898	80	0,217	0,172	0,854	21526	47	0,218	0,263	0,781	3547	6	0,169	0,246	0,858
DNA/hAT-Blackjack	3188449	14738	0,46223	87159	282	0,32355	6,60647E-11	0,999999991	36898	103	0,27915	1,29213E-08	0,999999991	21526	57	0,2648	2,32585E-06	0,999998695	3547	19	0,53566	0,7842	0,2908
DNA/hAT-Charlie	3188449	127020	3,984	87159	1899	2,179	0,000	1,000	36898	1058	2,867	0,000	1,000	21526	682	3,168	0,000	1,000	3547	118	3,327	0,023	0,982
DNA/hAT-hAT19	3188449	3988	0,12508	87159	52	0,05966	6,03549E-10	0,999999991	36898	31	0,08402	0,011405698	0,992757131	21526	21	0,09756	0,145786098	0,896943132	3547	5	0,14096	0,7139	0,4559
DNA/hAT-hAT5	3188449	21404	0,671	87159	243	0,279	0,000	1,000	36898	150	0,407	0,000	1,000	21526	116	0,539	0,008	0,994	3547	30	0,846	0,912	0,123
DNA/hAT-hATm	3188449	10907	0,34208	87159	222	0,25471	1,69768E-06	0,999998756	36898	111	0,30083	0,091445787	0,922980642	21526	75	0,34842	0,593487043	0,452184155	3547	17	0,47928	0,9322	0,1082
DNA/hAT-Tag1	3188449	11250	0,353	87159	124	0,142	0,000	1,000	36898	56	0,152	0,000	1,000	21526	32	0,149	0,000	1,000	3547	4	0,113	0,005	0,998
DNA/hAT-Tip100	3188449	145979	4,57837	87159	2489	2,8557	1,0694E-153	1	36898	1393	3,77527	1,19641E-14	1	21526	953	4,4272	0,147119073	0,860359567	3547	159	4,48266	0,4125	0,6187
DNA/Kolobok-Hydra	3188449	2865	0,090	87159	409	0,469	1,000	0,000	36898	96	0,260	1,000	0,000	21526	37	0,172	1,000	0,000	3547	1	0,028	0,173	0,959
DNA/Maverick	3188449	8167	0,25614	87159	197	0,22602	0,038096883	0,967404124	36898	88	0,2385	0,270106809	0,763777093	21526	58	0,26944	0,68169697	0,367131725	3547	11	0,31012	0,795	0,304
DNA/Merlin	3188449	2046	0,064	87159	37	0,042	0,004	0,997	36898	13	0,035	0,012	0,994	21526	5	0,023	0,006	0,998	3547	1	0,028	0,336	0,898
DNA/MuLE-MuDR	3188449	917	0,02876	87159	54	0,06196	0,999999901	2,26243E-07	36898	17	0,04607	0,976801363	0,041941771	21526	10	0,04646	0,949612157	0,096943096	3547	0	0	0,3603	1
DNA/P	3188449	6726	0,211	87159	152	0,174	0,008	0,994	36898	60	0,163	0,021	0,985	21526	37	0,172	0,117	0,911	3547	4	0,113	0,133	0,940
DNA/PIF-ISL2EU	3188449	1382	0,04334	87159	28	0,03213	0,958005629	0,960216238	36898	5	0,01355	0,001326385	0,999619247	21526	1	0,00465	0,000891768	0,999914238	3547	0	0	0,2147	1
DNA/PIF-Spy	3188449	14582	0,457	87159	230	0,264	0,000	1,000	36898	132	0,358	0,002	0,999	21526	79	0,367	0,024	0,981	3547	9	0,254	0,038	0,981
DNA/PiggyBac	3188449	5290	0,16591	87159	122	0,13997	0,028633665	0,976725285	36898	51	0,13822	0,103149647	0,919289247	21526	33	0,1533	0,363907124	0,698595968	3547	9	0,25374	0,9241	0,1408
DNA/Sola	3188449	7995	0,251	87159	179	0,205	0,003	0,998	36898	65	0,176	0,002	0,999	21526	45	0,209	0,121	0,905	3547	6	0,169	0,216	0,878
DNA/Sola-2	3188449	2932	0,09196	87159	58	0,06655	0,00529747	0,99633427	36898	21	0,05691	0,011555206	0,993302615	21526	17	0,07897	0,312021657	0,766377268	3547	3	0,08458	0,5888	0,6331
DNA/TcMar-Fot1	3188449	3291	0,103	87159	107	0,123	0,967	0,041	36898	46	0,125	0,912	0,115	21526	29	0,135	0,934	0,094	3547	10	0,282	0,999	0,004
DNA/TcMar-m44	3188449	788	0,02471	87159	52	0,05966	0,999999994	1,06035E-08	36898	18	0,04878	0,997376117	0,005749807	21526	8	0,03716	0,9097387	0,168163487	3547	0	0	0,4159	1
DNA/TcMar-Mariner	3188449	20902	0,656	87159	1078	1,237	1,000	0,000	36898	335	0,908	1,000	0,000	21526	204	0,948	1,000	0,000	3547	35	0,987	0,992	0,013
DNA/TcMar-Pogo	3188449	53692	1,68395	87159	788	0,90409	1,05159E-87	0,999999995	36898	448	1,21416	7,45388E-14	1	21526	291	1,35185	4,89743E-05	0,999961355	3547	48	1,35326	0,0676	0,9488
DNA/TcMar-Tc1	3188449	139366	4,371	87159	5164	5,925	1,000	0,000	36898	1980	5,366	1,000	0,000	21526	1122	5,212	1,000	0,000	3547	157	4,426	0,585	0,447
DNA/TcMar-Tc2	3188449	1325	0,04156	87159	17	0,0195	0,000259446	0,999885386	36898	8	0,02168	0,03071792	0,985594283	21526	6	0,02787	0,210818091	0,881834158	3547	0	0	0,2287	1
DNA/TcMar-Tc4	3188449	21816	0,684	87159	659	0,756	0,995	0,005	36898	262	0,710	0,740	0,281	21526	155	0,720	0,755	0,272	3547	18	0,507	0,117	0,922
DNA/TcMar-Tigger	3188449	71621	2,24626	87159	1065	1,2219	5,364E-113	0,999999993	36898	634	1,71825	4,34921E-13	0,999999998	21526	423	1,96507	0,002387133	0,997948043	3547	70	1,9735	0,1486	0,8771
DNA/Zator	3188449	16524	0,518	87159	220	0,252	0,000	1,000	36898	128	0,347	0,000	1,000	21526	93	0,432	0,040	0,968	3547	16	0,451	0,341	0,743
DNA/Zisupton	3188449	1953	0,06125	87159	40	0,04589	0,032588999	0,977044849	36898	15	0,04065	0,059838573	0,963868058	21526	10	0,04646	0,235153731	0,847037246	3547	2	0,05639	0,63	0,639
LINE/CR1	3188449	3223	0,101	87159	164	0,188	1,000	0,000	36898	56	0,152	0,998	0,002	21526	23	0,107	0,657	0,423	3547	3	0,085	0,518	0,695
LINE/CR1-Zenon	3188449	378	0,01186	87159	45	0,05163	0,99999999	2,91168E-16	36898	17	0,04607	0,999999311	2,98011E-06	21526	5	0,02323	0,955009232	0,115184166	3547	0	0	0,6565	1
LINE/Dong-R4	3188449	21921	0,688	87159	4158	4,771	1,000	0,000	36898	980	2,656	1,000	0,000	21526	222	1,031	1,000	0,000	3547	20	0,564	0,218	0,840
LINE/I	3188449	34972	1,09683	87159	1779	2,0411	0,999999998	6,0568E-131	36898	645	1,74806	1	5,08839E-29	21526	357	1,65846	1	8,79876E-14	3547	66	1,86073	1	4E-05
LINE/I-Nimb	3188449	2695	0,085	87159	499	0,573	1,000	0,000	36898	144	0,390	1,000	0,000	21526	58	0,269	1,000	0,000	3547	4	0,113	0,816	0,352
LINE/Jockey	3188449	5625	0,17642	87159	373	0,42795	0,999999997	1,42228E-52	36898	135	0,36587	1	1,5433E-14	21526	59	0,27409	0,999449947	0,000894702	3547	6	0,16916	0,5649	0,595
LINE/L1	3188449	1028	0,032	87159	42	0,048	0,995	0,008	36898	27	0,073	1,000	0,000	21526	15	0,070	0,998	0,005	3547	4	0,113	0,994	0,029
LINE/L1-Tx1	3188449	374	0,01173	87159	6	0,00688	0,113299027	0,943338937	36898	5	0,01355	0,732648589	0,435354133	21526	2	0,00929	0,536963787	0,718849153	3547	1	0,02813	0,9342	0,3405
LINE/L2	3188449	42322	1,327	87159	1520	1,744	1,000	0,000	36898	569	1,542	1,000	0,000	21526	322	1,496	0,985	0,018	3547	63	1,776	0,990	0,015
LINE/Penelope	3188449	62053	1,94618	87159	3029	3,47526	0,999999996	1,5711E-197	36898	1083	2,93512	1	3,07011E-38	21526	581	2,69906	0,999999998	1,8315E-14	3547	82	2,31181	0,9461	0,0676
LINE/R1	3188449	1976	0,062	87159	69	0,079	0,981	0,026	36898	28	0,076	0,880	0,164	21526	13	0,060	0,536	0,574	3547	1	0,028	0,355	0,889
LINE/R1-LOA	3188449	714	0,02239	87159	25	0,02868	0,911019808	0,128060151	36898	8	0,02168	0,555715353	0,58407658	21526	4	0,01858	0,472032971	0,709809782	3547	0	0	0,4517	1
LINE/RTE-BovB	3188449	20952	0,657	87159	707	0,811	1,000	0,000	36898	290	0,786	0,999	0,001	21526	176	0,818	0,998	0,003	3547	25	0,705	0,685	0,390
LINE/RTE-RTE	3188449	78216	2,4531	87159	19574	22,4578	0,999999998	0	36898	3810	10,3258	1	0	21526	816	3,79076	1	2,0539E-32	3547	82	2,31181	0,3167	0,7216
LINE/RTE-X	3188449	690	0,022	87159	14	0,016	0,153	0,899	36898	8	0,022	0,595	0,546	21526	6	0,028	0,811	0,324	3547	3	0,085	0,992	0,043
low_complexity	3188449	41405	1,29859	87159	422	0,48417	1,0413E-133	0,999999996	36898	314	0,85099	2,54411E-16	1	21526	229	1,06383	0,000941023	0,999244795	3547	54	1,52241	0,8926	0,1357
LTR	3188449	22969	0,720	87159	1288	1,478	1,000	0,000	36898	493	1,336	1,000	0,000	21526	238	1,106	1,000	0,000	3547	34	0,959	0,957	0,062
LTR/Copia	3188449	1775	0,05567	87159	77	0,08834	0,99993585	7,6868AE-05	36898	22	0,05962	0,678688784	0,40233763	21526	11	0,0511	0,4630774	0,651798823	3547	0	0	0,1386	1
LTR/ERV1	3188449	211	0,007	87159	17	0,020	1,000	0,000	36898	6	0,016	0,988	0,037	21526	2	0,009	0,828	0,417	3547	0	0,000	0,791	1,000
LTR/ERVK	3188449	7990	0,25059	87159	147	0,16866	1,2467E-07	0,999999912	36898	81	0,21952	0,124158589	0,896844982	21526	60	0,27873	0,816427104	0,220694024	3547	6	0,16916	0,2166	0,8778
LTR/Gypsy	3188449	7031	0,221	87159	467	0,536	1,000	0,000	36898	184	0,499	1,000	0,000	21526	86	0,400	1,000	0,000	3547	16	0,451	0,997	0,007
LTR/Pao	3188449	3390	0,10632	87159	530	0,60808	0,999999996	1,5949E-227	36898	170	0,46073	1	8,17136E-55	21526	56	0,26015	0,999999996	3,04683E-09	3547	12	0,33831	0,9998	0,0005
RC/Helitron	3188449	2444	0,077	87159	66	0,076	0,493	0,557	36898	42	0,114	0,994	0,009	21526	34	0,158	1,000	0,000	3547	4	0,113	0,860	0,290
rRNA	3188449	851	0,02669	87159	261	0,29945	0,999999994	1,1629E-189	36898	92	0,24934	1	1,75145E-57	21526	56	0,26015	0,999999995	3,50901E-36	3547	0	0	0,3878	1
Satellite	3188449	4890	0,153	87159	166	0,190	0,997	0,003	36898	76	0,206	0,995	0,008	21526	39	0,181	0,870	0,168	3547	6	0,169	0,696	0,461
SINE?	3188449	12984	0,40722	87159	401	0,46008	0,993233303	0,007812063	36898	243	0,65857	0,999999999	1,46766E-12	21526	148	0,68754	0,999999993	2,17254E-09	3547	22	0,62024	0,9775	0,0379
SINE/ID	3188449	143670	4,506	87159	2754	3,160	0,000	1,000	36898	1326	3,594	0,000	1,000	21526	918	4,265	0,044	0,959	3547	162	4,567	0,590	0,441
SINE/tRNA-Core	3188449	5423	0,17008	87159	80	0,09179	3,68199E-10	1	36898	55	0,14906	0,178902356	0,853640582	21526	37	0,17189	0,569054079	0,496529315	3547	9	0,25374	0,9141	0,1559
SINE/tRNA-Deu	3188449	14305	0,449	87159	189	0,217	0,000	1,000	36898	118	0,320	0,000	1,000	21526	82	0,381	0,072	0,942	3547	16	0,451	0,575	0,525
SINE/tRNA-Deu-CR1	3188449	10858	0,34054	87159	128	0,14686	3,73048E-29	1	36898	78	0,21139	2,86752E-06	0,99999827	21526	53	0,24621	0,007713463	0,994654093	3547	9	0,25374	0,235	0,8507
Unknown	3188449	1840236	57,716	87159	31928	36,632	0,000	1,000	36898	18022	48,843	0,000	1,000	21526	11847	55,036	0,000	1,000	3547	2004	56,498	0,073	0,931

[a]Results generated using Bedtools (analysis of the intersection of the annotation by repeatmasker and the HCR annotations). The number of hits for each element could slightly differ from those results reported in Table S1-9a, as in this analysis we have not take into account overlapping annotations, or multiple annotations in the same region.

[b]Total number of repeat annotations for the whole genome (n = 3,188,449). This value is slightly different from those reported by RepeatMasker (n = 3,284,969) . In the current analysis we have not take into account overlapping annotations, or multiple annotations in the same region.

[c]Total number of hits for every type of repeat class for the whole genome

[d]Total number of annotation for this subset

[e]Total number of hit for every type of repeat class for this subset

[f]CDF = P-value of observing (obtaining by random) a number of repeat elements less than or equal to the observed

[g]SF = P-value of observing (obtaining by random) a number of repeat elements greater than or equal to the observed

***Table S2-5*: Enrichment analysis of the intersection of High Coverage regions (HCRs) (5 x mean coverage) with RepeatMasker annotation (subtypes of main repeats)[a]**

				> 150 bp					> 500 bp					> 1000 bp					> 5000 bp				
	Total Genome[b]	Total Genome type[c]	%	Total Count Subset[d]	Total Subset Type[e]	%	CDF[f]	SF[g]	Total Count Subset	Total Subset Type	%	CDF	SF	Total Count Subset	Total Subset Type	%	CDF	SF	Total Count Subset	Total Subset Type	%	CDF	SF
DNA/CMC-Chapaev-3	3188449	5416	0,170	22089	28	0,127	0,065	0,955	8240	15	0,182	0,670	0,429	4106	10	0,244	0,904	0,167	392	0	0,000	0,514	1,000
DNA/CMC-EnSpm	3188449	10378	0,32549	22089	39	0,17656	1,47531E-05	0,999992252	8240	15	0,18204	0,009480618	0,995055159	4106	9	0,21919	0,14247753	0,916262668	392	0	0	0,2786	1
DNA/CMC-Transib	3188449	1359	0,043	22089	5	0,023	0,092	0,958	8240	1	0,012	0,134	0,970	4106	0	0,000	0,173	1,000	392	0	0,000	0,846	1,000
DNA/Crypton-H	3188449	1430	0,04485	22089	6	0,02716	0,135185158	0,930021458	8240	3	0,03641	0,494853581	0,714256235	4106	2	0,04871	0,71951717	0,549614786	392	0	0	0,8387	1
DNA/Ginger	3188449	9241	0,290	22089	19	0,086	0,000	1,000	8240	5	0,061	0,000	1,000	4106	1	0,024	0,000	1,000	392	1	0,255	0,686	0,679
DNA/hAT	3188449	364071	11,4184	22089	1483	6,71375	3,876E-124	0,999999998	8240	751	9,11408	6,25797E-12	0,999919996	4106	414	10,0828	0,003874031	0,997109321	392	45	11,4796	0,5547	0,5083
DNA/hAT-Ac	3188449	7724	0,242	22089	45	0,204	0,134	0,895	8240	17	0,206	0,299	0,777	4106	4	0,097	0,030	0,989	392	1	0,255	0,754	0,614
DNA/hAT-Blackjack	3188449	14738	0,46223	22089	55	0,24899	2,09331E-07	0,999999888	8240	23	0,27913	0,005793546	0,996694161	4106	11	0,2679	0,034625838	0,981733261	392	0	0	0,1626	1
DNA/hAT-Charlie	3188449	127020	3,984	22089	431	1,951	0,000	1,000	8240	228	2,767	0,000	1,000	4106	132	3,215	0,005	0,996	392	14	3,571	0,401	0,698
DNA/hAT-hAT19	3188449	3988	0,12508	22089	11	0,0498	0,000277036	0,999895861	8240	10	0,12136	0,544704156	0,580061989	4106	7	0,17048	0,852343085	0,257742234	392	0	0	0,6122	1
DNA/hAT-hAT5	3188449	21404	0,671	22089	78	0,353	0,000	1,000	8240	34	0,413	0,001	0,999	4106	16	0,390	0,012	0,993	392	0	0,000	0,071	1,000
DNA/hAT-hATm	3188449	10907	0,34208	22089	49	0,22183	0,000697413	0,999563784	8240	31	0,37621	0,740489516	0,32204425	4106	12	0,29226	0,353499367	0,744399973	392	1	0,2551	0,6122	0,739
DNA/hAT-Tag1	3188449	11250	0,353	22089	41	0,186	0,000	1,000	8240	17	0,206	0,011	0,994	4106	10	0,244	0,145	0,912	392	0	0,000	0,250	1,000
DNA/hAT-Tip100	3188449	145979	4,57837	22089	585	2,64838	2,31977E-50	1	8240	323	3,9199	0,001883805	0,99842347	4106	190	4,62737	0,578693776	0,450616667	392	24	6,12245	0,938	0,0937
DNA/Kolobok-Hydra	3188449	2865	0,090	22089	65	0,294	1,000	0,000	8240	10	0,121	0,871	0,212	4106	2	0,049	0,287	0,883	392	0	0,000	0,703	1,000
DNA/Maverick	3188449	8167	0,25614	22089	39	0,17656	0,00844164	0,99445601	8240	31	0,37621	0,984094392	0,025277457	4106	17	0,41403	0,978053801	0,039798146	392	2	0,5102	0,9192	0,2657
DNA/Merlin	3188449	2046	0,064	22089	9	0,041	0,101	0,944	8240	5	0,061	0,566	0,609	4106	0	0,000	0,072	1,000	392	0	0,000	0,778	1,000
DNA/MuLE-MuDR	3188449	917	0,02876	22089	16	0,07243	0,999686234	0,000874339	8240	6	0,07282	0,989235786	0,033665936	4106	5	0,12177	0,998633181	0,007204291	392	0	0	0,8934	1
DNA/P	3188449	6726	0,211	22089	33	0,149	0,023	0,985	8240	14	0,170	0,251	0,824	4106	7	0,170	0,365	0,761	392	1	0,255	0,799	0,563
DNA/PIF-ISL2EU	3188449	1382	0,04334	22089	2	0,00905	0,003831823	0,999285943	8240	0	0	0,027961105	1	4106	0	0	0,168429864	1	392	0	0	0,8437	1
DNA/PIF-Spy	3188449	14582	0,457	22089	54	0,244	0,000	1,000	8240	22	0,267	0,004	0,998	4106	13	0,317	0,106	0,934	392	2	0,510	0,733	0,536
DNA/PiggyBac	3188449	5290	0,16591	22089	32	0,14487	0,250197404	0,801554873	8240	15	0,18204	0,701744266	0,394674516	4106	8	0,19484	0,753433854	0,373303939	392	1	0,2551	0,8614	0,4785
DNA/Sola	3188449	7995	0,251	22089	44	0,199	0,067	0,950	8240	29	0,352	0,969	0,048	4106	10	0,244	0,546	0,579	392	1	0,255	0,742	0,626
DNA/Sola-2	3188449	2932	0,09196	22089	19	0,08602	0,442293659	0,645257069	8240	11	0,1335	0,916520205	0,144103211	4106	3	0,07306	0,478282243	0,722484613	392	1	0,2551	0,9488	0,3028
DNA/TcMar-Fot1	3188449	3291	0,103	22089	17	0,077	0,130	0,913	8240	8	0,097	0,522	0,615	4106	6	0,146	0,863	0,253	392	1	0,255	0,937	0,333
DNA/TcMar-m44	3188449	788	0,02471	22089	11	0,0498	0,989854067	0,023655609	8240	3	0,03641	0,850737962	0,333202454	4106	1	0,02435	0,730323097	0,637794974	392	0	0	0,9076	1
DNA/TcMar-Mariner	3188449	20902	0,656	22089	159	0,720	0,889	0,127	8240	60	0,728	0,814	0,224	4106	35	0,852	0,947	0,075	392	5	1,276	0,954	0,118
DNA/TcMar-Pogo	3188449	53692	1,68395	22089	170	0,76961	1,84998E-32	1	8240	95	1,15291	4,56463E-05	0,999969569	4106	53	1,29079	0,025230609	0,981612574	392	5	1,27551	0,3526	0,7898
DNA/TcMar-Tc1	3188449	139366	4,371	22089	970	4,391	0,567	0,446	8240	370	4,490	0,713	0,305	4106	204	4,968	0,970	0,035	392	17	4,337	0,551	0,548
DNA/TcMar-Tc2	3188449	1325	0,04156	22089	3	0,01358	0,018391335	0,994718361	8240	3	0,03641	0,552927719	0,665320127	4106	3	0,07306	0,906031693	0,244374443	392	0	0	0,8496	1
DNA/TcMar-Tc4	3188449	21816	0,684	22089	111	0,503	0,000	1,000	8240	47	0,570	0,115	0,910	4106	25	0,609	0,320	0,747	392	3	0,765	0,718	0,503
DNA/TcMar-Tigger	3188449	71621	2,24626	22089	309	1,39889	3,76277E-20	0,999999999	8240	158	1,91748	0,021868647	0,981911426	4106	93	2,26498	0,559650078	0,482048014	392	12	3,06122	0,892	0,1762
DNA/Zator	3188449	16524	0,518	22089	57	0,258	0,000	1,000	8240	37	0,449	0,215	0,829	4106	27	0,658	0,908	0,130	392	3	0,765	0,852	0,332
DNA/Zisupton	3188449	1953	0,06125	22089	7	0,03169	0,040281574	0,98141049	8240	4	0,04854	0,431960574	0,741987542	4106	2	0,04871	0,539840548	0,716064095	392	0	0	0,7865	1
LINE/CR1	3188449	3223	0,101	22089	35	0,158	0,995	0,008	8240	10	0,121	0,782	0,325	4106	7	0,170	0,940	0,127	392	0	0,000	0,673	1,000
LINE/CR1-Zenon	3188449	378	0,01186	22089	11	0,0498	0,999982321	8,4576E-05	8240	4	0,04854	0,996740628	0,017443075	4106	1	0,02435	0,913886948	0,385607239	392	0	0	0,9546	1
LINE/Dong-R4	3188449	21921	0,688	22089	1111	5,030	1,000	0,000	8240	156	1,893	1,000	0,000	4106	35	0,852	0,912	0,120	392	0	0,000	0,067	1,000
LINE/I	3188449	34972	1,09683	22089	485	2,19566	1	6,56172E-44	8240	157	1,90534	1	9,97995E-11	4106	85	2,07014	0,999999968	5,81639E-08	392	9	2,29592	0,9876	0,0309
LINE/I-Nimb	3188449	2695	0,085	22089	154	0,697	1,000	0,000	8240	34	0,413	1,000	0,000	4106	14	0,341	1,000	0,000	392	1	0,255	0,956	0,282
LINE/Jockey	3188449	5625	0,17642	22089	109	0,49346	1	1,79939E-20	8240	33	0,40049	0,999991211	2,10756E-05	4106	12	0,29226	0,966228413	0,06484011	392	1	0,2551	0,8472	0,4995
LINE/L1	3188449	1028	0,032	22089	15	0,068	0,997	0,006	8240	8	0,097	0,998	0,006	4106	6	0,146	1,000	0,002	392	0	0,000	0,881	1,000
LINE/L1-Tx1	3188449	374	0,01173	22089	2	0,00905	0,520164794	0,731964509	8240	2	0,02427	0,925997925	0,251916884	4106	2	0,04871	0,987043398	0,08457578	392	0	0	0,9551	1
LINE/L2	3188449	42322	1,327	22089	261	1,182	0,029	0,975	8240	113	1,371	0,659	0,376	4106	62	1,510	0,862	0,169	392	7	1,786	0,846	0,267
LINE/Penelope	3188449	62053	1,94618	22089	654	2,96075	1	1,3306E-24	8240	220	2,6699	0,999997425	3,6158E-06	4106	113	2,75207	0,999835318	0,000241286	392	8	2,04082	0,6447	0,4951
LINE/R1	3188449	1976	0,062	22089	9	0,041	0,124	0,929	8240	4	0,049	0,422	0,750	4106	1	0,024	0,278	0,922	392	0	0,000	0,784	1,000
LINE/R1-LOA	3188449	714	0,02239	22089	3	0,01358	0,271620507	0,871683518	8240	2	0,02427	0,718545269	0,550825729	4106	0	0	0,398452994	1	392	0	0	0,9159	1
LINE/RTE-BovB	3188449	20952	0,657	22089	139	0,629	0,322	0,708	8240	55	0,667	0,582	0,472	4106	32	0,779	0,856	0,189	392	2	0,510	0,524	0,729
LINE/RTE-RTE	3188449	78216	2,4531	22089	5475	24,7861	1	0	8240	672	8,15534	1	1,2446E-155	4106	111	2,70336	0,861125693	0,161672004	392	14	3,57143	0,9374	0,1065
LINE/RTE-X	3188449	690	0,022	22089	4	0,018	0,479	0,704	8240	4	0,049	0,965	0,106	4106	1	0,024	0,777	0,589	392	1	0,255	0,997	0,081
Low_complexity	3188449	41405	1,29859	22089	123	0,55684	2,88353E-28	1	8240	81	0,98301	0,004986284	0,996356056	4106	47	1,14467	0,213284838	0,826067363	392	7	1,78571	0,8584	0,2503
LTR	3188449	22969	0,720	22089	255	1,154	1,000	0,000	8240	72	0,874	0,953	0,060	4106	35	0,852	0,862	0,180	392	0	0,000	0,059	1,000
LTR/Copia	3188449	1775	0,05567	22089	11	0,0498	0,427555552	0,683965191	8240	7	0,08495	0,906386933	0,179986021	4106	4	0,09742	0,91808381	0,197638272	392	0	0	0,8039	1
LTR/ERV1	3188449	211	0,007	22089	0	0,000	0,231	1,000	8240	0	0,000	0,579	1,000	4106	0	0,000	0,762	1,000	392	0	0,000	0,974	1,000
LTR/ERVK	3188449	7990	0,25059	22089	16	0,07243	4,12864E-10	1	8240	7	0,08495	0,000490013	0,999844602	4106	6	0,14613	0,112472716	0,943425566	392	0	0	0,3474	1
LTR/Gypsy	3188449	7031	0,221	22089	113	0,512	1,000	0,000	8240	35	0,425	1,000	0,000	4106	13	0,317	0,924	0,128	392	0	0,000	0,421	1,000
LTR/Pao	3188449	3390	0,10632	22089	104	0,4708	1	9,51983E-35	8240	20	0,24272	0,999997603	0,000749463	4106	11	0,2679	0,998143111	0,005320013	392	0	0	0,659	1
RC/Helitron	3188449	2444	0,077	22089	11	0,050	0,086	0,950	8240	6	0,073	0,556	0,604	4106	5	0,122	0,901	0,210	392	0	0,000	0,740	1,000
rRNA	3188449	851	0,02669	22089	243	1,1001	0,999999999	2,6054E-308	8240	95	1,15291	0,999999997	1,1683E-119	4106	44	1,0716	0,999999992	1,98124E-54	392	0	0	0,9006	1
Satellite	3188449	4890	0,153	22089	107	0,484	1,000	0,000	8240	47	0,570	1,000	0,000	4106	19	0,463	1,000	0,000	392	1	0,255	0,878	0,452
SINE?	3188449	12984	0,40722	22089	182	0,8234	0,999999999	6,41903E-18	8240	89	1,0801	0,999999995	1,17641E-15	4106	40	0,97418	0,999999644	8,75723E-07	392	1	0,2551	0,5257	0,798
SINE/ID	3188449	143670	4,506	22089	656	2,970	0,000	1,000	8240	330	4,005	0,014	0,988	4106	180	4,384	0,371	0,658	392	17	4,337	0,499	0,598
SINE/tRNA-Core	3188449	5423	0,17008	22089	19	0,08602	0,000614837	0,999970445	8240	12	0,14563	0,356577486	0,741751591	4106	8	0,19484	0,731425548	0,398837982	392	1	0,2551	0,8558	0,4869
SINE/tRNA-Deu	3188449	14305	0,449	22089	58	0,263	0,000	1,000	8240	27	0,328	0,054	0,963	4106	16	0,390	0,338	0,746	392	2	0,510	0,742	0,525
SINE/tRNA-Deu-CR1	3188449	10858	0,34054	22089	33	0,1494	3,17019E-08	0,999999993	8240	13	0,15777	0,001212041	0,999471017	4106	7	0,17048	0,031635811	0,985784466	392	0	0	0,2626	1
Unknown	3188449	1840236	57,716	22089	8191	37,082	0,000	1,000	8240	4212	51,117	0,000	1,000	4106	2226	54,213	0,000	1,000	392	216	55,102	0,160	0,864

[a]Results generated using Bedtools (analysis of the intersection of the annotation by repeatmasker and the HCR annotations). The number of hits for each element could slightly differ from those results reported in Table S1-9a, as in this analysis we have not take into account overlapping annotations, or multiple annotations in the same region.

[b]Total number of repeat annotations for the whole genome (*n* = 3,188,449). This value is slightly different from those reported by RepeatMasker (*n* = 3,284,969) . In the current analysis we have not take into account overlapping annotations, or multiple annotations in the same region.

[c]Total number of hits for every type of repeat class for the whole genome

[d]Total number of annotation for this subset

[e]Total number of hit for every type of repeat class for this subset

[f]CDF = a *P*-value of observing (obtaining by random) a number of repeat elements less than or equal to the observed

[g]SF = a P-value of observing (obtaining by random) a number of repeat elements greater than or equal to the observed

Capítulo 5

Discusión

El desarrollo reciente y vertiginiso de las tecnologías de secuenciación ha posibilitado el acceso a la comunidad científica al uso y análisis de datos masivos. Han aparecido numerosas tecnologías basadas en la secuenciación tanto de nucleótidos (ADN y ARN) como también de proteínas que han permitido abordar cuestiones que resultaban inalcanzables por los métodos tradicionales. Además, el nivel de desarrollo es tan grande que se pueden realizar análisis de células individuales o de un conjunto e incluso a diferentes niveles a la vez: genético, transcriptional, regulatorio, epigenético, conformacional... El desarrollo de la bioinformática ha sido crucial en la aplicación de estas metodologías y en la interpretación de los resultados (Box 2). De igual forma, también ha sido muy relevante el papel de algunos proyectos genómicos de gran interés (Box 3) que por su envergadura, implicaciones y aplicabilidad han revolucionado el campo y la forma de hacer tanto los análisis como de su comprensión.

Hemos visto como el número de secuencias genómicas ha crecido notablemente en la última década en las bases de datos (Figura 2) y hoy en día representa una amplia diversidad de especies (Tabla 1). Los conocimientos aportados por los diferentes proyectos genómicos son múltiples y con importante trascendencia en la ciencia tanto básica como aplicada (Tablas 2-3) [5, 78, 92–95]. Pero a pesar de que encontramos una representación similar entre los diferentes reinos y dentro de cada subgrupo, existen muchas familias de organismos sin apenas representación en las bases de datos. Este sesgo se puede deber por una parte, a las limitaciones

del uso extendido y rutinario de tanto la bioinformática como de la secuenciación [108]. Por otra parte, existe una parte de la comunidad científica, principalmente la que estudia organismos no modelo, que no son capaces de realizar este tipo de aproximaciones usando datos masivos. Prefieren seguir usando un conjunto de marcadores de los que disponen información previa.

Los organismos de estudio de esta tesis doctoral han sido organismos no modelo pertenecientes al grupo de los artrópodos, y en concreto al grupo de los arácnidos (Figura 6). Las arañas presentan aspectos biológico-evolutivos muy interesantes (Box 7) pero la escasa representación de estos organismos en las bases de datos genómicos [138] podría estar ralentizando el potencial descubrimiento de nuevos recursos y su repercusión tanto a nivel de la biología básica, como de la aplicada. Por ejemplo, la disponibilidad de datos genómicos, transcriptómicos y proteómicos son indispensables para poder caracterizar los genes y proteínas que componen los diferentes tipos de fibras de seda y de venenos. Por una parte, el desarrollo de aplicaciones biotecnológicas de las fibras de seda mediante ingeniería genética precisa de las secuencias genómicas codificantes de estas fibras, de muy alta calidad y continuidad [170]. Por otra parte, en cuanto a los venenos, debido a su diversidad y complejidad molecular es necesaria información tanto de transcriptomas como de proteomas para poder caracterizar bien estos compuestos biologicamente activos [141]. Además, la disponibilidad de recursos genómicos de buena calidad permite determinar si la diversidad molecular de estos venenos es debida a familias multigénicas, a genes parálogos, a diferencias entre alelos, a mecanismos de *splicing* alternativo, entre otros mecanismos. Por otro lado, la disponibilidad de datos masivos o la capacidad de obtener marcadores moleculares de una forma rápida y efectiva, ayudará a entender las relaciones filogenéticas entre los arácnidos que tantas controversias ha creado a lo largo de la historia [134–136, 168, 169].

5.1. Implementación de nuevos métodos bioinformáticos para el desarrollo y búsqueda de marcadores moleculares.

Los análisis filogenéticos a partir de pocos marcadores [116] o sólo basados en marcadores mitocondriales [117] presentan múltiples desventajas y limitaciones. Para muchos estudios es necesario emplear varios marcadores, no ligados y repartidos aleatoriamente a lo largo del genoma [118, 119]. Las tecnologías de NGS permiten de forma eficiente y rápida la obtención de una gran cantidad de información genómica, pero el principal inconveniente está en el procesamiento bioinformático y selección de los marcadores de interés.

Existen diversas herramientas bioinformáticas para el desarrollo de marcadores moleculares que usan estrategias de partición genómicas basadas en NGS (Tabla 4). Mientras que algún tipo de marcadores como EPIC [122] y NPCL [120, 121] dependen de la presencia de genomas filogenéticamente cercanos anotados, otros como los UCEs [115] no presentan este requisito. *A priori* se podría obtener marcadores moleculares de cualquier organismo aplicando la técnica de UCEs, pero puede ser limitante en muchos casos ya que es necesario la disponibilidad de un conjunto de secuencias obtenido previamente de un organismo relativamente relacionado. La obtención de las regiones UCEs se produce principalmente por técnicas NGS basadas en métodos de enriquecimiento y captura de secuencias. Resulta imprescindible por tanto el disponer de un conjunto de secuencias a capturar, específicamente diseñadas, que incluya el rango taxonómico de nuestra especie de interés. También encontramos por otra parte, los marcadores asociados a RRL como RADseq [112, 113] y que son ampliamente utilizados. La principal limitación de esta técnica radica en el propio diseño ya que tan solo unas pocas bases adyacentes al sitio de restricción son secuenciadas y por tanto, en general, no se pueden realizar análisis de marcadores ligados. Además, se puede producir la pérdida de las diferentes dianas de restricción por divergencia interespecífica y disminuir su posibilidad de uso en un rango taxonómico amplio.

Para facilitar el desarrollo de marcadores moleculares hemos desarrollado la aplicación DOMINO: *Development of Molecular Markers in Non Model organisms* [171]. Se trata de una herramienta bioinformática que facilita la identificación

y selección de marcadores moleculares a partir de una serie de características de interés proporcionadas por el usuario, como por ejemplo la longitud deseada para el marcador, el nivel de variabilidad mínimo necesario, el rango taxonómico de interés, etc. Permite generar marcadores adecuados para estudios a nivel filogenético, filogeográfico o a nivel de genética de poblaciones [94, 119, 126, 172]. DOMINO permite desarrollar regiones conservadas y variables que se puedan utilizar (i) directamente como marcadores moleculares con una profundidad taxonómica determinada, (ii) para la amplificación por PCR de en una rango taxonómico más o menos alejado y (iii) para el desarrollo de regiones para métodos de captura de secuencia. La herramienta está diseñada para aumentar la flexibilidad en el desarrollo de marcadores y facilitar el uso por parte de investigadores, especialmente en organismos no modelo. Presenta una versión *command-line* y una multiplataforma con una interfaz gráfica (GUI, del inglés *"Graphical User Interface"*) (Figura 9).

DOMINO utiliza datos de NGS, o alineamientos múltiples (MSA) pre-computados. Aunque inicialmente fue desarrollada para datos de secuenciación de la plataforma *Roche 454* no apareados (SE, del inglés *"Single end"*), pronto se vio la necesidad de adaptarla a los nuevos tipos de datos, *Illumina* SE pero también apareados (PE). *A posteriori* y para aprovechar las capacidades y características del diseño de la herramienta se decidió ampliar a MSA pre-computados, aumentando la flexibilidad y versatilidad de la herramienta. Además, DOMINO puede generar ensamblajes *de novo* o utilizar un genoma de referencia disponible de una especie cercana que se quiera incluir en el análisis.

Una estrategia fácil y asequible, que se puede aplicar para la generación de marcadores moleculares utilizando DOMINO en un amplio rango taxonómico, es el diseño de marcadores moleculares anónimos (ANMs) mediante RRLs. Los marcadores moleculares generados por esta metodología tendrán una distribución uniforme, aleatoria, no sesgada a una región y sometidos a evolución neutra [107, 125, 126]. Aunque este tipo de estrategia mediante RRLs comparte limitaciones con RADseq, por las sustituciones nucleotídicas en las dianas de restricción, la metodología es diferente y presenta ciertas ventajas. Tras la fragmentación del genoma se selecciona por tamaño tan sólo aquellos fragmentos más largos y se genera la secuenciación de estos. Se obtiene una reducción genómica que abarata los costes de secuenciación y permite generar estudios de ligamiento de los marcadores obtenidos, dentro del mismo fragmento secuenciado. Una vez obtenida

la reducción genómica de una selección de taxones representativos del rango filogenético a estudiar, se puede utilizar DOMINO para diseñar marcadores que satisfagan diferentes características (longitud de las zonas variables y conservadas, rango de variabilidad, taxones incluidos...). Utilizando las regiones conservadas de los marcadores obtenidos se podrán generar *primers* para amplificar por PCR o métodos de captura basados en NGS en el resto de taxones dentro del rango filogenético de interés. Empleando esta metodología estaremos generando múltiples marcadores, no ligados y repartidos aleatoriamente a lo largo del genoma a partir de individuos cuya información genómica previa no se conocía.

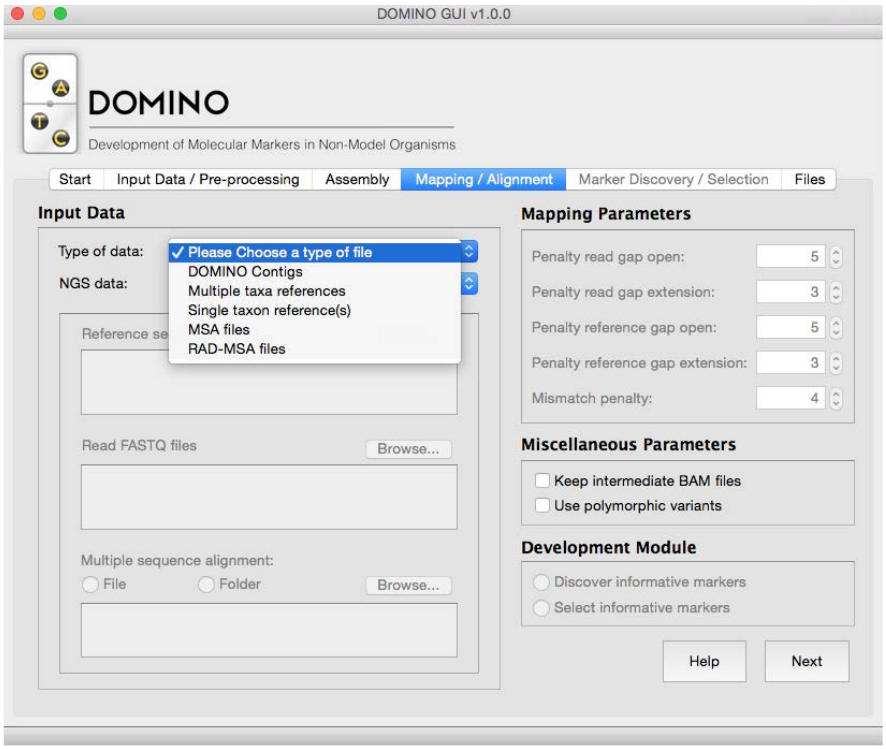


Figura 9: **Interfaz gráfica de DOMINO**. Fuente [171].

Pero DOMINO también acepta otro tipo de datos. Aunque originalmente

fue diseñada para el desarrollo de estos marcadores ANMs obtenidos mediante RRLs y a partir de *Roche 454* ó *Illumina* (SE o PE), se amplió su versatilidad para poder maximizar la utilidad de la herramienta. En el método anterior descrito, uno de los módulos de la herramienta alinea o mapea la información NGS frente a una secuencia de referencia, ya sea un genoma que proporciona el usuario de una especie cercana o a partir de ensamblaje *de novo* de las secuencias proporcionadas. A partir de este mapeo a una referencia se produce un alineamiento de las diferentes secuencias de las especies proporcionadas para generar un MSA para cada uno de los regiones compartidas entre taxones. A partir de aquí, otro módulo se encargará de comprobar la idoneidad de estos MSAs dadas unas características de interés. Es por tanto posible, acceder a este punto del proceso aportando MSAs pre-computados de algunos marcadores de interés (*e.g.* genes individuales) o a partir de estudios masivos como RADseq. Mediante esta entrada de datos pre-computados se selecciona sólo aquellas regiones que cumplen con las características deseadas por el usuario. De esta forma, se puede maximizar las capacidades de procesamiento de la información que proporciona DOMINO empleando datos pre-computados y de diferente origen.

La aplicación de DOMINO ha sido validada mediante simulaciones computacionales y datos empíricos. Las simulaciones *in silico* de datos NGS, con perfiles de error mimetizando el error producido por las tecnologias de secuenciación (específico de cada plataforma), permitieron ajustar y configurar los parámetros para maximizar la sensibilidad y precisión en la búsqueda e identificación de los marcadores. A partir de una topología fija de árbol filogenético de 4 taxones, se simularon experimentos RRL donde se generaron 100 fragmentos de entre 3-10 kb con diferente tasa de sustituciones por sitio (0.1-0.3) y diferente heterogeneidad a lo largo de las regiones. A partir de estas secuencias, se simularon *reads* empleando perfiles de calidad y tasas de error de secuencias reales para diferentes tipos de tecnologías NGS (*Roche 454*, *Illumina* SE y PE). Además, estos *reads* se simularon con diferentes longitudes y a diferentes niveles de profundidad de secuenciación (5x-20x). Para cada combinación de parámetros se generaron 500 réplicas y se analizó la sensibilidad y la precisión para recuperar marcadores moleculares obtenidos de las secuencias originales. La combinación de los resultados de las diferentes simulaciones generadas permitió conocer y ajustar los parámetros en cada etapa del proceso.

Por otra parte, se realizó una validación empírica con un conjunto de datos reales. Se seleccionaron 4 individuos pertenecientes a 3 especies del género *Nemesia*, (Nemesiidae, Mygalomorphae, Araneae) (Figura 6) a partir de los cuales se realizó una secuenciación mediante la tecnología *Roche 454* tras una reducción genómica con RRLs. Utilizando la herramienta DOMINO se realizó el control de calidad, ensamblaje de los *reads* y búsqueda de marcadores. Con los resultados obtenidos se comprobó la idoneidad de algunos de estos marcadores para recuperar la filogenia de los individuos de interés. Además se confirmó su validez para resolver la filogenia de otras especies, taxonómicamente relacionadas, mediante amplificación por PCR de las regiones conservadas de algunos marcadores. Esta validación empírica permitió demostrar un ejemplo de uso de esta herramienta bioinformática en el campo de la filogenia de arácnidos mediante el uso de RRLs y marcadores ANMs.

DOMINO tiene un diseño modular y versátil. Por una parte, resulta muy fácil la implementación de nuevos módulos o herramientas para acomodar nuevos datos de secuenciación (*e.g.* tecnologías de *reads* largos) o a partir de nuevos alineamientos o datos pre-computados. Además, presenta la posibilidad de ejecutarse tanto por la *command-line* como la GUI (Figura 9) lo que hace que esta herramienta pueda ser empleada tanto en entornos de alta computación como por usuarios menos familiarizados con los entornos de programación.

Esta herramienta está orientada a investigadores que deseen obtener marcadores moleculares empleando técnicas NGS con unos conocimientos medios o bajos de entornos de programación. La versatilidad y la posibilidad de definir diferentes características para los marcadores de interés, hacen que DOMINO puede ser aplicable a un amplio rango de aproximaciones desde la filogenia, genética de poblaciones y filogeografía. La interfaz hace de DOMINO una herramienta de fácil uso y accesible a la comunidad científica para el manejo de datos de NGS en la identificación y desarrollo de marcadores moleculares en organismos no modelo, principalmente.

Esta herramienta DOMINO es una aplicación de código libre y abierto, liberada bajo los términos de la *"Free Software Foundation/GPL License"*. Está disponible en el repositorio de versiones *Github* (https://github.com/molevol-ub/DOMINO) y en la página web del grupo de investigación donde se ha desarrollado esta tesis doctoral (http://www.ub.edu/softevol/domino/).

5.2. Ensamblaje genómico y de alta calidad de un representante del género *Dysdera*.

Existen diferentes proyectos internacionales con una intención firme y clara de secuenciar y promover buenas prácticas entre los investigadores que desarrollan proyectos genómicos en eucariotas. Por una parte existe un proyecto a nivel del grupo de artrópodos conocido como i5k (del inglés *"5000 Arthropod Genome Initiative"*) [173], lanzado en 2012. Por otra parte, encontramos otro de reciente creación (2018) y conocido como *"Earth BioGenome Project"* (EBP) [174] y que representa a un rango más amplio de especies y que incluye a todos los eucariotas. Ambas iniciativas ponen de manifiesto la importancia de secuenciar un número elevado de especies tanto para su conocimiento biológico básico como por su potencial en la conservación y preservación de la biodiversidad y en aplicaciones biotecnológicas, industriales, biomédicas, agriculturales o de control de plagas, etc. Como ya hemos visto, son múltiples y muy valiosos los conocimientos aportados por los proyectos de secuenciación genómica. Se han podido llevar a cabo estudios de procesos evolutivos y de evolución molecular y se han abordado cuestiones evolutivas, de conservación o ecológicas sin precedentes [94, 107]. Además, se han podido identificar genes o familias de genes específicos de linaje con mayor o menor repercusión fenotípica [78].

Los organismos de estudio de esta tesis doctoral, las arañas (Figura 6), son un grupo de depredadores que ocupan un amplio abanico de nichos ecológicos y con más de 45.000 especies documentadas [128], están especialmente infrarrepresentadas en las bases de datos genómicos (Tabla 5). Apenas encontramos 6 secuencias genómicas completas de arañas disponibles, la mayoría representantes del grupo Entelegynae y tan solo un representante del grupo Synspermiata y otro del gran orden Mygalomorphae. En un estudio reciente [138] se identificó las familias y grandes grupos de arañas sin representación genómica pero con importante relevancia para el conocimiento de las adaptaciones de este grupo de organismos (Box 7). Se identificó como potenciales grupos a aquellos con posiciones filogenéticas clave o que mantuvieran características ancestrales. Entre las diferentes familias nombradas por su potencial interés se encuentra la familia Dysderidae por su relevancia dentro del grupo Synspermiata.

Tabla 5: **Estadísticas de ensamblajes genómicos de arañas disponibles en *GenBank*.** Datos obtenidos a fecha de Septiembre 2019.

Grupo	Especie	Tamaño[a]	N50 (bp)	Genes	Año
Synspermiata	*Dysdera silvatica*	1,7	174.193.557[b]	48.619	2019[c]
Synspermiata	*Loxosceles reclusa*	3,7	63.237	20.617	2015[d]
Entelegynae	*Stegodyphus mimosarum*	2,7	480.636	27.252	2014[e]
Entelegynae	*Parasteatoda tepidariorum*	1,2	765.179	27.990[b]	2017[f]
Entelegynae	*Latrodectus hesperus*	1,2	39.474	17.364[b]	2015[d]
Entelegynae	*Nephila clavipes*	2,4	62.959	14.025	2017[h]
Mygalomorphae	*Acanthoscurria geniculata*	7,2	20.294	-	2014[e]

[a]Tamaño genómico (Gb). [b]Datos actualizados respecto a la version publicada referenciada. [c]Sánchez-Herrero et al. [175]; [d]i5k Consortium [173]; [e]Sanggaard et al. [139]; [f]Schwager et al. [140]; [h]Babb et al. [176];

El género *Dysdera*, principal género de la familia Dysderidae y representativo de ésta, presenta un total de 282 especies [128] y es el principal grupo de organismos que concierne a esta tesis doctoral. Estas arañas, de hábitos nocturnos [129, 130] y con una distribución principalmente en la cuenca mediterránea y las Islas Macaronésicas [131], son conocidas por la increíble radiación adaptativa del género en las Islas Canarias (Figura 8) [155–157]. Estas arañas ocupan una variedad muy grande de nichos ecológicos y existen unas 55 especies endémicas de las islas, con ejemplos de endemismos dentro de cada isla [156–159]. Otro aspecto destacable de este género es que son un ejemplo de estenofagia en arañas, presentan una preferencia de dieta hacia los isópodos [160, 161]. Esta característica no ocurre en todas las especies de *Dysdera* y encontramos especies especialistas, con dietas basadas principalmente en isópodos y otras especies generalistas de dieta, con un abanico mucho más amplio que puede incluir o no a isópodos. Encontramos una diversificación de la forma y tamaño de los quelíceros que se corresponde con la preferencia de presa hacia los isópodos [160, 161]. Esta preferencia implica adaptaciones tanto metabólicas y morfológicas (quelíceros) como de comportamiento (estrategias de captura) [160, 161, 164–167] (Box 8).

Puesto que las bases moleculares tanto de la radiación adaptativa del género como de la especialización trófica son completamente desconocidas y dado el interés de generar nuevos datos genómicos en el grupo de las arañas y en concreto de Synspermiata, procedimos a generar un ensamblaje y anotación funcional de un representante de este género en este caso de *Dysdera silvatica*. Esta especie en concreto fue seleccionada por presentar un tipo de dieta generalista y ser endémica de las Islas Canarias.

En el ensamblaje de *D. silvatica* se emplearon hasta 5 tipos de tecnologías NGS diferentes. La combinación de metodologías complementarias permitió una mejora notable de la calidad del ensamblaje (Figura 10). Se puede apreciar cómo tras la adición de cada librería de secuenciación se consigue una mejora en las diferentes estadísticas, ya sea de continuidad (N50) o del grado de integridad del genoma (en inglés *"completeness"*) medido como el porcentaje de secuencias encontradas de un conjunto de genes de copia única y compartidos entre todos los artrópodos, utilizando el *software* BUSCO (del inglés *"Benchmarking of Universal Single Copy Orthologs"*) [177]. Se puede apreciar también la inherente inversión económica y la repercusión sobre el resultado final.

Inicialmente, se generó una secuenciación masiva mediante la plataforma *Illumina* PE (100 PE; ~300 pb inserto) a partir del material genético de un individuo de *D. silvatica*, y la secuenciación mediante *Illumina* MP (100 PE, 5 kb inserto), de otro individuo, ambas con una cobertura media razonable dado el tamaño genómico estimado por citometría de flujo (~1,7 Gb). Esta librería MP podría *a priori* saltar las zonas repetitivas y poder generar *scaffold*s de los *contig*s ensamblados. Lamentablemente, una vez adicionada esta nueva librería la calidad del genoma apenas se vió mejorada. Vimos la necesidad de generar una nueva secuenciación debido a la complejidad del mismo. Se probaron diferentes estrategias para intentar reducir la variabilidad introducida (polimorfismo) por el material genético de dos individuos pero los resultados seguían siendo pobres y lejos de un ensamblaje de calidad. Debido a la popularización de los métodos de secuenciación de fragmentos largos durante el transcurso de la tesis se produjo una reducción de costes y fue asequible el poder realizar este tipo de aproximaciones. En la figura 10 se puede apreciar como el resultado de añadir una primera librería de *PacBio* (5 *lanes*) no tuvo una gran repercusión en la calidad, sobre todo por una cuestión de cobertura. Posteriormente hicimos una segunda librería, en este caso de *Nanopore*, que generó un incremento notable tanto en calidad como en

la continuidad. Con esta librería, casi un 90 % de los genes compartidos entre los artrópodos de copia única (genes BUSCO) eran recuperados a partir de este ensamblaje. Además, la continuidad del genoma (N50) estaba en torno a 38 kb valores similares al resto de ensamblajes de las especies de arañas disponibles en las base de datos *GenBank* (Tabla 5). En cada uno de los pasos de adición de librerías de secuenciación, la elección de los *software* de ensamblaje así como la optimización de los parámetros fue hecha en base a bibliografía, teniendo en cuenta la disponibilidad de las librerías de secuenciación en cada caso, y a estadísticas del ensamblaje tanto de la continuidad como de la integridad mediante *scripts* bioinformáticos desarrollados durante la tesis doctoral.

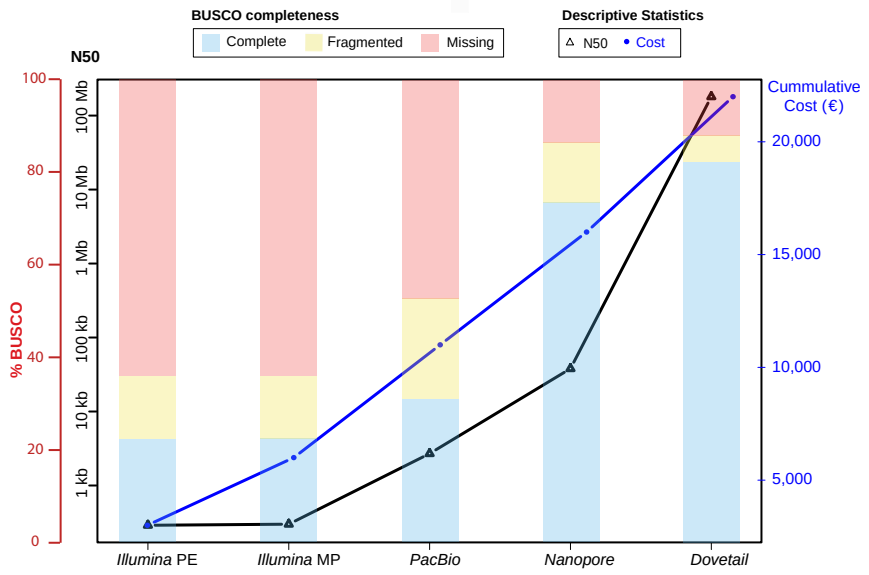


Figura 10: **Estadísticas del ensamblaje de *D. silvatica* tras la adición de diferentes librerias de secuenciación**. Las barras indican los porcentajes de cada categoría de resultados de BUSCO (completo, fragmentado y ausente, en diferentes colores) utilizando los genes ortólogos 1:1 del conjunto de artrópodos. En negro, en escala logarítmica, el valor del estadístico N50 (pb). En azul el coste acumulado (€). Ver detalle en tabla S5.

A partir de este ensamblaje se produjo una anotación estructural y funcional. Se emplearon métodos de predicciones génicas *ab initio* que mediante iteraciones permitieron el entrenamiento de los diferentes modelos. Además, se emplearon métodos de predicción de genes basados en evidencias de transcritos (datos de RNAseq de la propia *D. silvatica* [145]) o de proteínas bien anotadas y curadas en las diferentes bases de datos correspondientes a representantes de los grandes linajes de artrópodos. La integración de las diferentes fuentes de información permitió optimizar y maximizar la identificación de regiones codificantes y no codificantes en cada caso. Una vez realizada la anotación de genes, se procedió a determinar la posible función de estos mediante la inferencia con bases de datos de dominios anotados de familias conocidas. Del total de proteínas identificadas (48.619), en torno a un 75 % (36.398) presentaban dominios proteicos o similitud de secuencia alta con proteínas bien anotadas en las bases de datos. Este trabajo dio lugar a la publicación del primer genoma representativo de la familia Dysderidae y superfamilia Dysderoidea, segundo del grupo Synspermiata y apenas el séptimo de todo el orden Araneae [175].

La disponibilidad de este ensamblaje genómico supone un punto de partida para el análisis de características biológicas específicas del grupo pero también del análisis de peculiaridades compartidas con otros arácnidos o artrópodos. Por una parte, se podrán estudiar en profundidad características genómico-evolutivas del grupo de las arañas (Box 7), como el desarrollo del veneno y de la seda. También podrá aportar conocimiento sobre la duplicación genómica ocurrida en el grupo de las arañas tras la diversificación de los escorpiones [140] o los eventos de terrestralización de los artrópodos. Además, permitirá profundizar en el conocimiento del sistema quimiosensorial de arañas descrito previamente [145]. De igual forma, puesto que se trata de una familia sin previa representación y con un interés alto por su posicionamiento filogenético [138] tendrá una gran relevancia en los análisis filogenéticos y aportará conocimiento a las discrepancias en torno a la filogenia de las arañas y quelicerados [134–136, 168, 169].

Por otra parte, la disponibilidad de este recurso será vital para estudios futuros sobre las bases moleculares de la radiación adaptativa del género *Dysdera* en la Islas Canarias. De igual forma, se podrá estudiar el proceso de especialización de la dieta de este grupo y los mecanismos moleculares subyacentes a los cambios metabólicos y morfológicos necesarios. Además, como este género presenta casos de adaptación a ambientes hipógeos, ha desarrollado complejos órganos repro-

ductivos y mecanismos de selección femenica críptica; la disponibilidad de este genoma puede aportar también un conocimiento muy valioso en este sentido.

Puesto que esta tesis se engloba dentro de un proyecto de investigación más amplio donde se están analizando las bases moleculares de la adaptación mediante aproximaciones genómico-comparativas y de genómica de poblaciones de este género *Dysdera*, se procedió a utilizar este ensamblaje genómico como base para mejorar su continuidad para futuros análisis. Se procedió a la generación de una librería de captura de la conformación cromosómica denominada Hi-C y realizada por la empresa *Dovetail Genomics*. Esta nueva librería y su posterior análisis bioinformático permitieron aumentar la continuidad del genoma en 4 órdenes de magnitud pasando de un valor de N50 de 38 kb a casi ~175 Mb (Figura 10; Tabla S5) (Ref: Datos no publicados del grupo de investigación). Se puede apreciar como la integridad del genoma, es decir, el número de genes recuperados de los compartidos con otros artrópodos no cambia al añadir esta nueva librería pero si mejora su continuidad al reducirse el número de genes fragmentados. El hecho de que sigan apareciendo en torno a un 10 % de genes BUSCO ausentes no tiene porque ser debido a que este genoma está algo incompleto, sino del conjunto de datos empleado para verificarlo. Para la generación del conjunto de genes de copia única compartidos entre todos los artrópodos, los autores del *software* BUSCO, Simao *et al.* [177], apenas utilizaron representantes de quelicerados por lo que probablemente existe un importante sesgo. A día de hoy, no existe ningún otro conjunto de genes compartido de copia única más específico para el grupo de las arañas que el de los artrópodos.

Los resultados preliminares son claros y muestran que este ensamblaje de calidad cromosómica, donde practicamente cada cromosoma viene representado por un único *scaffold*, supondría a día de hoy, el ensamblaje más continuo del grupo de las arañas disponibles en *GenBank* (Tabla 5) y mejor que el de muchos otros genomas disponibles para otros grupos de organismos. La adición de librerías de secuenciación de fragmentos largos así como la de librerías de captura de la conformación cromosómica ha sido esencial para conseguir el alto nivel de calidad, continuidad e integridad del genoma de *D. silvatica*.

5.3. Repercusión de las técnicas de secuenciación masivas y bioinformática en organismos no modelo.

A pesar de que el número de genomas disponibles en las bases de datos ha crecido considerablemente en los últimos años (Tabla 1), y de la reducción de costes de la secuenciación masiva (Figura 2), no ha sido rápido, directo ni rutinario generar un ensamblaje de buena calidad ni obtener marcadores moleculares.

La reducción de costes de la secuenciación masiva ha sido notable en los últimos años pero siguen existiendo limitaciones inherentes de las propias técnicas. Algunas de estas restricciones se han ido mitigando y hoy en día, la combinación de datos de secuenciación generados por diferentes metodologías es una buena solución, especialmente en organismos con genomas grandes y complejos. También existen limitaciones en el análisis e interpretación de los datos, sobre todo para una parte de la comunidad científica, menos versada en el uso de herramientas bioinformáticas.

Para generar un ensamblaje genómico de alta calidad, sobre todo de un genoma grande y complejo como el utilizado en esta tesis doctoral, es necesario invertir una cantidad importante de dinero en la generación de datos de diferentes técnicas aunque los resultados son notorios y satisfactorios (Figura 10). Por otra parte, en el proceso de desarrollo de marcadores moleculares, donde la generación de datos resulta asequible, se pone de manifiesto la necesidad de la bioinformática para la identificación y selección de aquellas regiones más relevantes.

El desarrollo de la bioinformática paralelo al desarrollo de la secuenciación masiva ha permitido la interpretación y el análisis de los resultados y su aplicación en numerosos campos de la biología. El análisis de los resultados es complejo ya que precisa de personal formado tanto con conocimientos de la biología como de conocimientos en bioinformática. Por una parte los conocimientos biológicos permitirán una interpreación de los resultados mientras que los conocimientos mas técnicos de bioinformática permitirán utilizar herramientas o análisis de una forma más eficiente, rapida y precisa. A la vez, se necesita de datos de confianza para poder comprobar la calidad de las predicciones y ajustar los parámetros de los diferentes *softwares*. La bioinformática ha facilitado la interpretación de

los resultados al proveer al usuario final de métodos para la visualización de la información, y al facilitar el uso de las diferentes herramientas por usuarios menos experimentados con los entornos de programación.

La escasez de recursos genómicos y la limitación del análisis de datos masivos se ve acrecentada en estudios en organismos no modelo. Pero es, precisamente, en éstos, donde hoy en día es muy importante y potencialmente valioso el poder generar y disponer de tanto un genoma completo y anotado, como de información para su uso en el desarrollo de marcadores moleculares. Las posibilidades que presentan el disponer de un nuevo genoma ensamblado de calidad cromosómica son muy importantes y van desde análisis detallados de características biológicas específicas de la especie de interés, a análisis comparativos del género, familia o cualquier otro rango taxonómico.

5

Capítulo 6

Conclusiones

6

1. La herramienta bionformática desarrollada en este tesis, a la que hemos llamado DOMINO, permite la identificación y selección de marcadores moleculares en organismos no modelo a partir de datos de secuenciación masiva o alineamientos múltiples pre-computados.

2. DOMINO es capaz de usar datos de diferentes plataformas de secuenciación masiva e indicar características de interés para los marcadores diseñados, pudiendose aplicar de esta manera a estudios filogenéticos, filogeográficos o de genética de poblaciones. Los resultados de la herramienta pueden emplearse de forma directa como marcadores, ser usados para su amplificación por PCR en un conjunto de especies con un rango taxonómico deseado o utilizarlos en el desarrollo de sondas para métodos de obtención de marcadores por captura de secuencia.

3. La validación del software mediante simulaciones computacionales, bajo diferentes condiciones de longitud de secuencia y divergencia, mimetizando los perfiles de error de secuenciación de datos reales, ha permitido ajustar y configurar los parámetros que maximizan la sensibilidad y precisión de nuestra herramienta.

4. Mediante el análisis de datos empíricos, hemos demostrado la aplicabilidad de DOMINO para generar marcadores moleculares anónimos en especie no modelo a partir de datos de secuenciación masiva de una librería genómica reducida usando enzimas de restricción.

5. La interfaz gráfica implementada en DOMINO permite el facil uso de la herramienta por parte de usuarios poco familiarizados con el tratamiento bioinformático de datos de secuenciación masiva y sus entornos de programación.

6. El draft de la secuencia genómica completa de un representante del género *Dysdera*, obtenido mediante la combinación de diferentes librerías de secuenciación y una estrategia de ensamblaje híbrida, supone el primer recurso genómico completo de un representante de la familia Dysderidae, además del segundo del grupo Synspermiata y tan solo el séptimo dentro de todo el orden Araneae.

7. El uso tanto de predicciones *ab initio* como de evidencias de diferentes bases de datos genómicas y de datos de RNAseq ha permitdio una anotación estructural y funcional del genóma de *D. silvatica* de alta calidad.

8. La complejidad del genoma de *D. silvatica*, reflejado en un alto número de regiones repetitivas muy intercaladas, ha limitado la continuidad del ensamblaje (N50 de 38 kb), pero no ha afectado la integridad de los genes, la cual presenta unas buenas estadísticas comparado con genomas de otras especies cercanas.

9. La calidad del genoma ensamblado, medida mediante diferentes estadísticas descriptivas y conjuntos de datos de RNAseq de la propia especie y de genes conservados a diferentes rangos filogenéticos, es alta especialmente en cuanto a su continuidad e integridad de los genes anotados, especialmente en relación a otros genomas de arañas.

10. Dada la poca representación de arañas en las bases de datos genómicas, el genoma anotado de *D. silvatica* supone un recurso enormemente útil para muchas áreas de la biología evolutiva y aplicada, siendo un punto de partida necesario para el análisis de características específicas del género pero también del análisis de peculiaridades compartidas con otros arácnidos o artrópodos.

Capítulo 7

Tablas Suplementarias

A continuación se incluyen tablas suplementarias donde se encuentra la información que se ha recopilado, de diferentes fuentes, para la elaboración de algunas ilustraciones de esta tesis doctoral.

7

Tabla S1: **Hitos en la historia de la secuenciación.** Fuente [4, 5]

Year	Category	Milestone
1953	Technical	DNA structure
1953	Technical	Insulin protein sequence
1965	Technical	Alanine tRNA sequence
1968	Application	Cohesive ends *Bacteriophage Lambda*
1973	Application	*Lac* operator binding site sequence
1977	Application	First genome sequencing
1977	Genome	*Bacteriophage X174*
1977	Technical	Maxam-Gilbert Sequencing Method
1977	Technical	Sanger Sequencing Method
1980	Computational	Protein Sequence Database (PSD)
1981	Application	Shotgun sequencing
1981	Computational	Smith-Waterman
1982	Computational	GenBank
1982	Genome	Bacteriophage lambda
1983	Application	Expressed sequence tags (EST)
1984	Computational	Protein Identification Resource (PIR)
1985	Technical	Polymerase chain reaction (PCR)
1986	Technical	Fluorescent detection in sequencing
1988	Technical	Stepwise dNTP incorporation
1990	Application	HGP initiation
1990	Computational	BLAST
1990	Technical	PE sequencing
1993	Technical	Optical mapping
1995	Application	Serial analysis of gene expression
1995	Genome	*Haemophilus influenzae*
1996	Computational	RepeatMasker
1996	Genome	*Saccharomyces cerevisae*
1996	Technical	Pyrosequencing
1997	Computational	GENSCAN
1998	Application	Human SNP discovery
1998	Genome	*Caenorhabditis elegans*
2000	Computational	Celera assembler
2000	Genome	*Drosophila melanogaster*
2000	Genome	*Arabidopsis thaliana*
2000	Technical	Sequencing by ligation (SBL)
2001	Computational	Bioconductor
2001	Genome	*Homo sapiens*
2002	Computational	UCSC Genome Browser
2002	Computational	Ensembl
2002	Genome	*Mus musculus*

Continuación en la siguiente pág.

Tabla S1 – *Continuación de la anterior pág.*

Año	**Tipo**	**Hito**
2003	Computational	UniProt
2003	Technical	Emulsion PCR
2003	Technical	Sequencing by synthesis (SBS)
2004	Application	Metagenome assembly (NGS)
2004	Genome	*Rattus norvegicus*
2005	Application	Bacterial assembly
2005	Computational	Galaxy
2005	Genome	*Pan troglodytes*
2005	Genome	*Oryza sativa*
2005	Technical	Nucleotide Reversible Terminators (NRTs)
2007	Application	ChIP-seq
2007	Application	First cancer genome sequencing
2007	Computational	NCBI SRA
2007	Technical	Targeted sequenced capture
2008	Application	RNA-seq
2008	Application	ATAC-seq
2008	Computational	ALL-PATHS
2009	Application	Exome sequencing
2009	Application	Ribosome profiling
2009	Application	scRNA-seq
2009	Computational	SAMtools, Bowtie, Tophat
2009	Genome	*Zea mays*
2010	Application	1000 Genomes Project (Phase I)
2010	Application	Large genome assembly from short reads
2010	Computational	SOAPdenovo
2010	Genome	Giant Panda (*Ailuropoda melanoleura*)
2010	Genome	Neanderthal
2010	Technical	Single-based resolution detector
2011	Application	Haplotype-resolved human genome
2011	Computational	Integrative Genomics Viewer (IGV)
2012	Genome	Denisovan genome sequencing
2012	Technical	Nanopore sequencing
2012	Technical	Single-stranded library for ancient DNA
2013	Genome	*Danio rerio*
2015	Application	1000 Genomes Project (Completition)
2016	Application	Pacbio Human genome sequencing
2016	Application	ONT Human genome sequencing
2017	Computational	CANU assembler
2017	Genome	*Xenopus laevis*

Tabla S2: **Cronología de la historia de la secuenciación:** se indica la cantidad de bases (pb) depositads en *GenBank* [9]; el número de artículos en *PubMed* [10] que contienen *"Bioinformatics"*, *"Next Generation Sequencing"* [NGS] y *"Genome Sequencing"* [GS]; y el coste ($/pb) [7].

Año	**pb**	**Bioinformatics**	**GS**	**NGS**	**Coste**
1982	680.338	0	0	0	0
1983	2.274.029	0	0	0	0
1984	3.689.752	1	0	0	0
1985	5.204.420	0	0	0	0
1986	9.615.371	3	257	0	0
1987	16.752.872	10	334	0	0
1988	24.690.876	11	455	0	0
1989	37.183.950	60	601	1	0
1990	51.306.092	125	676	0	0
1991	77.337.678	196	898	0	0
1992	120.242.234	163	995	1	0
1993	163.802.597	181	1.190	4	0
1994	230.485.928	166	1.362	2	0
1995	425.860.958	197	1.432	1	0
1996	730.552.938	304	1.527	2	0
1997	1.258.290.513	346	1.663	7	0
1998	2.162.067.871	682	1.736	6	0
1999	4.653.932.745	834	1.759	13	0
2000	1.1101.066.288	1.505	2.007	15	0
2001	15.849.921.438	2.237	1.983	26	5.292,39
2002	28.507.990.166	3.288	2.059	29	3.898,64
2003	36.553.368.485	4.690	1.995	22	2.230,98
2004	44.575.745.176	6.432	2.313	36	1.028,85
2005	56.037.734.462	8.196	2.392	37	766,73
2006	69.019.290.705	8.911	2.454	31	581,92
2007	83.874.179.730	9.990	2.517	60	397,09
2008	99.116.431.942	10.919	2.641	126	3,81
2009	110.118.557.163	12.100	2.987	305	0,78
2010	122.082.812.719	13.497	3.732	696	0,32
2011	135.117.731.375	15.115	4.568	1.435	0,09
2012	148.390.863.904	16.673	5.403	2.243	0,07
2013	156.230.531.562	17.845	6.970	3.457	0,06
2014	184.938.063.614	22.421	8.743	4.802	0,06
2015	203.939.111.071	25.647	9.856	6.148	0,014
2016	224.973.060.433	28.010	11.262	7.291	0,015
2017	249.722.163.594	29.478	11.179	7.485	0,012
2018	285.688.542.186	28.669	11.081	7.417	0,012
2019	321.680.566.570	117.40	4.069	2.633	0,012

Tabla S3: **Datos de rendimiento de las principales metodologías de secuenciación.** Fuente: [43, 44]

Tecnología	**Plataforma**	**Año**	***Reads*[a]**	**Longitud[b]**	**Rendimiento[c]**
ABI Sanger	3.730xl	2002	96	800	0,0000768
Roche 454	GS20	2005	200.000	100	0,02
Roche 454	GS FLX+	2011	1.000.000	700	0,7
Illumina	GA	2006	28.000.000	25	0,7
Illumina	GAIIx	2011	640.000.000	75	48
Illumina	HiSeq 2000/2500	2014	4.000.000.000	125	1.000
Illumina	HiSeq X	2014	6.000.000.000	150	1.800
SOLiD	1	2007	40.000.000	25	1
SOLiD	5500xl	2011	3.000.000.000	60	180
SOLiD	5500xl W	2013	3.000.000.000	75	320
Ion Torrent	PGM 314 chip	2011	100.000	100	0.01
Ion Torrent	Ion S5/S5XL 540 chip	2015	75.000.000	200	15
PacBio	RS C1	2011	432.000	1.300	0,54
PacBio	RS II P6 C4	2014	660.000	13500	12
Oxford Nanopore	MinION Mk1 fast	2015	4.400.000	9.545	42

[a] *Reads* por *run* de secuenciación
[b] Longitud media
[c] Bases por *run* (Gb)

Tabla S4: **Distribución de las especies de *Dysdera* en la Islas Canarias** (H: El Hierro, P: La Palma; T: Tenerife; C: Gran Canaria; F: Fuerteventura; L: Lanzarote). Fuente: [157]

Autor	**Año**	**Especie**	**H**	**P**	**G**	**T**	**C**	**F**	**L**
Wunderlich	1992	*Dysdera alegranzaensis*							X
Ribera, Ferrández and Blasco	1985	*Dysdera ambulotenta*				X			
Arnedo and Ribera	1997	*Dysdera andamanae*					X		
Arnedo and Ribera	1997	*Dysdera arabisenen*					X		
Schmidt	1973	*Dysdera bandamae*					X		
Wunderlich	1992	*Dysdera brevisetae*				X			
Wunderlich	1992	*Dysdera brevispina*				X			
Wunderlich	1987	*Dysdera calderensis*		X	X				
Wunderlich	1992	*Dysdera chioensis*				X			
Simon	1883	*Dysdera cribellata*				X			
Koch	1838	*Dysdera crocata*	X	X	X	X	X		
Wunderlich	1992	*Dysdera curvisetae*				X			
Arnedo, Oromí and Ribera	1997	*Dysdera enghoffi*			X				
Ribera and Blasco	1986	*Dysdera esquiveli*				X			
Wunderlich	1992	*Dysdera gibbifera*				X			
Ribera and Arnedo	1994	*Dysdera gollumi*				X			
Strand	1911	*Dysdera gomerensis*	X		X				
Arnedo and Ribera	1999	*Dysdera guayota*			X	X			
Arnedo and Ribera	1999	*Dysdera hernandezi*				X			
Arnedo, Oromí and Ribera	1997	*Dysdera hirguan*			X				
Wunderlich	1987	*Dysdera iguanensis*				X	X		
Simon	1883	*Dysdera insulana*				X	X		
Wunderlich	1992	*Dysdera labradaensis*				X			
Simon	1907	*Dysdera lancerotensis*						X	X
Wunderlich	1987	*Dysdera levipes*			X	X	X		
Simon	1907	*Dysdera liostethus*					X		
Wunderlich	1992	*Dysdera longa*						X	
Simon	1883	*Dysdera macra*				X			
Arnedo	2007	*Dysdera madai*				X			
Macías-Hernández and Arnedo	2010	*Dysdera mahan*						X	X
Wunderlich	1992	*Dysdera minutissima*				X			
Wunderlich	1992	*Dysdera montanetensis*				X			
Simon	1907	*Dysdera nesiotes*							X

Continuación en la siguiente pág.

Tabla S4 – *Continuación de la anterior pág.*

Autor	Año	Especie	H	P	G	T	C	F	L
Arnedo, Oromí and Ribera	1997	*Dysdera orahan*	X		X				
Wunderlich	1992	*Dysdera paucispinosa*					X		
Arnedo, Oromí and Ribera	1997	*Dysdera ramblae*			X				
Wunderlich	1992	*Dysdera ratonensis*		X					
Simon	1907	*Dysdera rugichelis*					X		
Arnedo, Oromí and Ribera	2000	*Dysdera sanborondon*						X	
Arnedo	2007	*Dysdera sibyllina*				X			
Schmidt	1981	*Dysdera silvatica*	X	X	X				
Macías-Hernández and Arnedo	2010	*Dysdera simbeque*							X
Wunderlich	1992	*Dysdera spinidorsum*						X	
Wunderlich	1992	*Dysdera tilosensis*					X		
Ribera, Ferrández and Blasco	1985	*Dysdera unguimmanis*				X			
Simon	1883	*Dysdera verneaui*				X			
Ribera, Ferrández and Blasco	1985	*Dysdera volcania*				X			
Arnedo and Ribera	1997	*Dysdera yguanirae*					X		

Tabla S5: **Estadísticas de las diferentes versiones del ensamblaje del genoma de *Dysdera silvatica*.**

Librería	Año	N50 (pb)	Coste[a]	Acumulado[a]	C[b]	F[c]	M[d]
Illumina PE	2015	290	3.000	3.000	22,2	13,5	64,3
Illumina MP	2016	300	3.000	6.000	22,2	13,5	64,3
PacBio	2017	2.700	5.000	11.000	30,9	21.8	47,3
ONT	2018	38.000	5.000	16.000	73,4	13	13,6
Dovetail Genomics	2019	174.193.557	6.000	22.000	82,2	5,8	12

[a] Expresado en Euros (€).
[b] Genes BUSCO completos (C) (%).
[c] Genes BUSCO fragmentedos (F) (%).
[d] Genes BUSCO ausentes (M) (%).

Capítulo 8

Bibliografía

Bibliografía

1. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics 2016 jan;107(1):1–8. http://www.ncbi.nlm.nih.gov/pubmed/26554401.
2. Gauthier J, Vincent AT, Charette SJ, Derome N. A brief history of bioinformatics. Briefings in Bioinformatics 2018 aug;https://doi.org/10.1093/bib/bby063.
3. Kircher M, Kelso J. High-throughput DNA sequencing - concepts and limitations. BioEssays 2010 may;32(6):524–536. http://www.ncbi.nlm.nih.gov/pubmed/20486139.
4. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics 2016;17(6):333–351. https://doi.org/10.1038/nrg.2016.49.
5. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, et al. DNA sequencing at 40: past, present and future. Nature 2017 oct;550(7676):345–353. https://doi.org/10.1038/nature24286.
6. Metzker ML. Sequencing technologies - the next generation. Nature reviews Genetics 2010 jan;11(1):31–46. http://www.ncbi.nlm.nih.gov/pubmed/19997069.
7. Wetterstrand KA, DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program (GSP); 2019. www.genome.gov/Sequencingcostsdata.
8. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Research 2017 nov;45(D1):D37–D42. https://doi.org/10.1093/nar/gks1195.
9. GenBank N, NCBI Genbank statistics; 2019. https://www.ncbi.nlm.nih.gov/genbank/statistics/.
10. PubMed N, NCBI PubMed statistics; 2019. https://www.ncbi.nlm.nih.gov/pubmed/.
11. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 1953 apr;171(4356):737–8. http://www.ncbi.nlm.nih.gov/pubmed/13054692.
12. Zallen DT. Despite Franklin's work, Wilkins earned his Nobel. Nature 2003 sep;425(6953):15–15. https://doi.org/10.1038/425015b.
13. Sanger F. Sequences, Sequences, and Sequences. Annual Review of Biochemistry 1988 jun;57(1):1–29. https://doi.org/10.1146/annurev.bi.57.070188.000245.

14. Wu R, Kaiser AD. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. Journal of Molecular Biology 1968 jan;35(3):523–537. https://doi.org/10.1016/S0022-2836(68)80012-9.
15. Gilbert W, Maxam A. The nucleotide sequence of the lac operator. Proceedings of the National Academy of Sciences of the United States of America 1973 dec;70(12):3581–4. http://www.ncbi.nlm.nih.gov/pubmed/4587255.
16. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America 1977 dec;74(12):5463–7. http://www.ncbi.nlm.nih.gov/pubmed/271968.
17. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, et al. Nucleotide sequence of bacteriophage ϕX174 DNA. Nature 1977 feb;265(5596):687–695. http://www.nature.com/doifinder/10.1038/265687a0.
18. Maxam AM, Gilbert W. A new method for sequencing DNA. Proceedings of the National Academy of Sciences 1977 feb;74(2):560–564. https://doi.org/10.1073/pnas.74.2.560.
19. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, et al. Fluorescence detection in automated DNA sequence analysis. Nature 1986 jun;321(6071):674–679. http://www.nature.com/doifinder/10.1038/321674a0.
20. Connell C, Fung S, Heiner C, Bridgham J, Chakerian V, Heron E, et al. Automated DNA-sequence analysis. BioTechniques 1987;5(4):342.
21. Saiki R, Scharf S, Faloona F, Mullis K, Horn G, Erlich H, et al. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science 1985 dec;230(4732):1350–1354. http://www.sciencemag.org/cgi/doi/10.1126/science.2999980.
22. Saiki R, Gelfand D, Stoffel S, Scharf S, Higuchi R, Horn G, et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 1988 jan;239(4839):487–491. http://www.sciencemag.org/cgi/doi/10.1126/science.2448875.
23. Jackson DA, Symons RH, Berg P. Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of Escherichia coli. Proceedings of the National Academy of Sciences of the United States of America 1972 oct;69(10):2904–9. http://www.ncbi.nlm.nih.gov/pubmed/4342968.
24. Cohen SN, Chang AC, Boyer HW, Helling RB. Construction of biologically functional bacterial plasmids in vitro. Proceedings of the National Academy of Sciences of the United States of America 1973 nov;70(11):3240–4. http://www.ncbi.nlm.nih.gov/pubmed/4594039.
25. Anderson S. Shotgun DNA sequencing using cloned DNase I-generated fragments. Nucleic Acids Research 1981 jul;9(13):3015–3027. https://doi.org/10.1093/nar/9.13.3015.
26. Staden R. A strategy of DNA sequencing employing computer programs. Nucleic Acids Research 1979 jun;6(7):2601–2610. https://doi.org/10.1093/nar/6.7.2601.

27. Klenow H, Henningsen I. Selective elimination of the exonuclease activity of the deoxyribonucleic acid polymerase from Escherichia coli B by limited proteolysis. Proceedings of the National Academy of Sciences of the United States of America 1970 jan;65(1):168–75. http://www.ncbi.nlm.nih.gov/pubmed/4905667.
28. Chen CY. DNA polymerases drive DNA sequencing-by-synthesis technologies: both past and present. Frontiers in Microbiology 2014 jun;5:305. http://www.ncbi.nlm.nih.gov/pubmed/25009536.
29. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005 sep;437(7057):376–380. http://www.nature.com/articles/nature03959.
30. Sutton GG, White O, Adams MD, Kerlavage AR. TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects. Genome Science and Technology 1995 jan;1(1):9–19. https://doi.org/10.1089/gst.1995.1.9.
31. Gordon D, Abajian C, Green P. Consed: a graphical tool for sequence finishing. Genome research 1998 mar;8(3):195–202. http://www.ncbi.nlm.nih.gov/pubmed/9521923.
32. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, et al. A whole-genome assembly of Drosophila. Science (New York, NY) 2000 mar;287(5461):2196–204. http://www.ncbi.nlm.nih.gov/pubmed/10731133.
33. Masic I, The most influential scientists in the development of medical informatics (13): Margaret Belle Dayhoff. The Academy of Medical Sciences of Bosnia and Herzegovina; 2016. http://www.ncbi.nlm.nih.gov/pubmed/27708497.
34. IUPAC-IUB Comm on Biochem Nomencl. A one-letter notation for amino acid sequences. Tentative rules. Biochemistry 1968 aug;7(8):2703–2705. http://www.ncbi.nlm.nih.gov/pubmed/5666745.
35. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 1970 mar;48(3):443–453. https://doi.org/10.1016/0022-2836(70)90057-4.
36. Feng DF, Doolittle RF. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. Journal of molecular evolution 1987;25(4):351–60. http://www.ncbi.nlm.nih.gov/pubmed/3118049.
37. Karsch-Mizrachi I, Takagi T, Cochrane G, International Nucleotide Sequence Database Collaboration. The international nucleotide sequence database collaboration. Nucleic acids research 2018 jan;46(D1):D48–D51. https://doi.org/10.1093/nar/gkx1097.
38. Wu CH, Yeh LSL, Huang H, Arminski L, Castro-Alvear J, Chen Y, et al. The Protein Information Resource. Nucleic acids research 2003 jan;31(1):345–7. https://doi.org/10.1093/nar/gkg040.
39. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. UniProt: the Universal Protein knowledgebase. Nucleic Acids Research 2004 jan;32(90001):115D–119. https://doi.org/10.1093/nar/gkh131.
40. Kanehisa M, Bork P, Bioinformatics in the post-sequence era; 2003. https://doi.org/10.1038/ng1109.
41. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology 1990 oct;215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

42. Brown SM. Get your bioinformatics on the Web! BioTechniques 2000 feb;28(2):244–6. http://www.ncbi.nlm.nih.gov/pubmed/10683733.
43. Nederbragt L, Developments in next generation sequencing. Figshare; 2016. https://figshare.com/articles/developments_in_NGS/100940/9.
44. Nederbragt L, Developments in next generation sequencing. Github; 2016. https://github.com/lexnederbragt/developments-in-next-generation-sequencing.
45. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature 2001 feb;409(6822):860–921. https://doi.org/10.1038/35057062.
46. Stein LD. Human genome: End of the beginning. Nature 2004 oct;431(7011):915–916. https://doi.org/10.1038/431915a.
47. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 2012 sep;489(7414):57–74. https://doi.org/10.1038/nature11247.
48. Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, et al. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science (New York, NY) 1987 oct;238(4825):336–41. http://www.ncbi.nlm.nih.gov/pubmed/2443975.
49. Tabor S, Richardson CC. DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. Proceedings of the National Academy of Sciences of the United States of America 1987 jul;84(14):4767–71. http://www.ncbi.nlm.nih.gov/pubmed/3474623.
50. Craxton M. Linear amplification sequencing, a powerful method for sequencing DNA. Methods 1991 aug;3(1):20–26. https://doi.org/10.1016/S1046-2023(05)80159-8.
51. DeAngelis MM, Wang DG, Hawkins TL. Solid-phase reversible immobilization for the isolation of PCR products. Nucleic acids research 1995 nov;23(22):4742–3. http://www.ncbi.nlm.nih.gov/pubmed/8524672.
52. Zhang JZ, Fang Y, Hou JY, Ren HJ, Jiang R, Roos P, et al. Use of Non-Cross-Linked Polyacrylamide for Four-Color DNA Sequencing by Capillary Electrophoresis Separation of Fragments up to 640 Bases in Length in Two Hours. Analytical Chemistry 1995 dec;67(24):4589–4593. https://doi.org/10.1021/ac00120a026.
53. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science (New York, NY) 2001 feb;291(5507):1304–51. http://www.ncbi.nlm.nih.gov/pubmed/11181995.
54. Tomkinson AE, Vijayakumar S, Pascal JM, Ellenberger T. DNA Ligases: Structure, Reaction Mechanism, and Function. Chemical Reviews 2006 feb;106(2):687–699. http://pubs.acs.org/doi/abs/10.1021/cr040498d.
55. Leamon JH, Lee WL, Tartaro KR, Lanza JR, Sarkis GJ, DeWinter AD, et al. A massively parallel PicoTiterPlate^{TM} based platform for discrete picoliter-scale polymerase chain reactions. ELECTROPHORESIS 2003 nov;24(21):3769–3777. http://www.ncbi.nlm.nih.gov/pubmed/14613204.
56. Guo J, Xu N, Li Z, Zhang S, Wu J, Kim DH, et al. Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. Proceedings of the National Academy of Sciences 2008 jul;105(27):9145–9150. http://www.ncbi.nlm.nih.gov/pubmed/18591653.

57. Wu J, Zhang S, Meng Q, Cao H, Li Z, Li X, et al. 3'-O-modified nucleotides as reversible terminators for pyrosequencing. Proceedings of the National Academy of Sciences of the United States of America 2007 oct;104(42):16462–7. https://doi.org/10.1073/pnas.0707495104{%}0A.
58. Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proceedings of the National Academy of Sciences 2003 jul;100(15):8817–8822. http://www.ncbi.nlm.nih.gov/pubmed/12857956.
59. Kim JB, Porreca GJ, Song L, Greenway SC, Gorham JM, Church GM, et al. Polony Multiplex Analysis of Gene Expression (PMAGE) in Mouse Hypertrophic Cardiomyopathy. Science 2007 jun;316(5830):1481–1484. http://www.ncbi.nlm.nih.gov/pubmed/17556586.
60. Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Research 2006 feb;34(3):e22–e22. https://doi.org/10.1093/nar/gnj023.
61. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, et al. Single-Molecule DNA Sequencing of a Viral Genome. Science 2008 apr;320(5872):106–109. http://www.sciencemag.org/cgi/doi/10.1126/science.1150427.
62. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, et al. Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. Science 2010 jan;327(5961):78–81. https://doi.org/10.1126/science.1181498.
63. Schatz MC, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing. Genome Research 2010 sep;20(9):1165–1173. http://www.ncbi.nlm.nih.gov/pubmed/20508146.
64. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-Time DNA Sequencing from Single Polymerase Molecules. Science 2009 jan;323(5910):133–138. http://www.ncbi.nlm.nih.gov/pubmed/19023044.
65. Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing. Nature Biotechnology 2016 may;34(5):518–524. http://www.ncbi.nlm.nih.gov/pubmed/27153285.
66. Bayley H. Nanopore Sequencing: From Imagination to Reality. Clinical Chemistry 2015 jan;61(1):25–31. https://doi.org/10.1373/clinchem.2014.223016.
67. Roberts RJ, Carneiro MO, Schatz MC. The advantages of SMRT sequencing. Genome Biology 2013 jul;14(7):405. https://doi.org/10.1186/gb-2013-14-7-405.
68. Li G, Cai L, Chang H, Hong P, Zhou Q, Kulakova EV, et al. Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application. BMC Genomics 2014 dec;15(S12):S11. https://doi.org/10.1186/1471-2164-15-S12-S11.
69. Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome research 2016 mar;26(3):342–50. http://www.ncbi.nlm.nih.gov/pubmed/26848124.

70. Moll KM, Zhou P, Ramaraj T, Fajardo D, Devitt NP, Sadowsky MJ, et al. Strategies for optimizing BioNano and Dovetail explored through a second reference quality assembly for the legume model, Medicago truncatula. BMC genomics 2017;18(1):578. http://www.ncbi.nlm.nih.gov/pubmed/28778149.
71. Neely RK, Deen J, Hofkens J. Optical mapping of DNA: Single-molecule-based methods for mapping genomes. Biopolymers 2011 may;95(5):298–311. http://doi.wiley.com/10.1002/bip.21579.
72. Schwartz DC, Li X, Hernandez LI, Ramnarain SP, Huff EJ, Wang YK. Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. Science (New York, NY) 1993 oct;262(5130):110–4. http://www.ncbi.nlm.nih.gov/pubmed/8211116.
73. Zhou S, Wei F, Nguyen J, Bechner M, Potamousis K, Goldstein S, et al. A Single Molecule Scaffold for the Maize Genome. PLoS Genetics 2009 nov;5(11):e1000711. https://dx.plos.org/10.1371/journal.pgen.1000711.
74. Teague B, Waterman MS, Goldstein S, Potamousis K, Zhou S, Reslewic S, et al. High-resolution human genome structure by single-molecule analysis. Proceedings of the National Academy of Sciences of the United States of America 2010 jun;107(24):10848–53. https://doi.org/10.1073/pnas.0914638107.
75. Nguyen QH, Pervolarakis N, Blake K, Ma D, Davis RT, James N, et al. Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. Nature Communications 2018 dec;9(1):2028. http://www.ncbi.nlm.nih.gov/pubmed/29795293.
76. López-Escardó D, Grau-Bové X, Guillaumet-Adkins A, Gut M, Sieracki ME, Ruiz-Trillo I. Evaluation of single-cell genomics to address evolutionary questions using three SAGs of the choanoflagellate Monosiga brevicollis. Scientific Reports 2017 dec;7(1):11025. http://www.nature.com/articles/s41598-017-11466-9.
77. Arendt D, Musser JM, Baker CVH, Bergman A, Cepko C, Erwin DH, et al. The origin and evolution of cell types. Nature Reviews Genetics 2016 dec;17(12):744–757. http://www.nature.com/articles/nrg.2016.127.
78. Ellegren H. Genome sequencing and population genomics in non-model organisms. Trends in ecology & evolution 2014 jan;29(1):51–63. http://www.ncbi.nlm.nih.gov/pubmed/24139972.
79. Genome N, NCBI Genome statistics; 2019. https://www.ncbi.nlm.nih.gov/genome/browse#!/eukaryotes/.
80. Edwards A, Voss H, Rice P, Civitello A, Stegemann J, Schwager C, et al. Automated DNA sequencing of the human HPRT locus. Genomics 1990 apr;6(4):593–608. http://www.ncbi.nlm.nih.gov/pubmed/2341149.
81. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 2009 jan;10(1):57–63. http://www.nature.com/articles/nrg2484.
82. Park PJ. ChIP–seq: advantages and challenges of a maturing technology. Nature Reviews Genetics 2009 oct;10(10):669–680. http://www.ncbi.nlm.nih.gov/pubmed/19736561.

83. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature Methods 2013 dec;10(12):1213–1218. http://www.ncbi.nlm.nih.gov/pubmed/24097267.
84. Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics (Oxford, England) 2013 nov;29(21):2669–77. http://www.ncbi.nlm.nih.gov/pubmed/23990416.
85. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences of the United States of America 2011 jan;108(4):1513–8. https://doi.org/10.1073/pnas.1017351108.
86. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nature reviews Genetics 2012 may;13(5):329–42. http://www.ncbi.nlm.nih.gov/pubmed/22510764.
87. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 genes. Science (New York, NY) 1996 oct;274(5287):546, 563–7. http://www.ncbi.nlm.nih.gov/pubmed/8849441.
88. C elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science (New York, NY) 1998 dec;282(5396):2012–8. http://www.ncbi.nlm.nih.gov/pubmed/9851916.
89. Initiative TAG. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 2000 dec;408(6814):796–815. http://www.nature.com/articles/35048692.
90. Adams MD. The Genome Sequence of Drosophila melanogaster. Science 2000 mar;287(5461):2185–2195. http://www.sciencemag.org/cgi/doi/10.1126/science.287.5461.2185.
91. Consortium MGS. Initial sequencing and comparative analysis of the mouse genome. Nature 2002 dec;420(6915):520–562. http://www.nature.com/articles/nature01262.
92. Losos JB, Arnold SJ, Bejerano G, Brodie ED, Hibbett D, Hoekstra HE, et al. Evolutionary Biology for the 21st Century. PLoS Biology 2013 jan;11(1):e1001466. http://dx.plos.org/10.1371/journal.pbio.1001466.
93. Vizueta J, Rozas J, Sánchez-Gracia A. Comparative Genomics Reveals Thousands of Novel Chemosensory Genes and Massive Changes in Chemoreceptor Repertories across Chelicerates. Genome Biology and Evolution 2018 may;10(5):1221–1236. https://doi.org/10.1093/gbe/evy081.
94. Lemmon EM, Lemmon AR. High-Throughput Genomic Data in Systematics and Phylogenetics. Annual Review of Ecology, Evolution, and Systematics 2013 nov;44(1):99–121. https://doi.org/10.1146/annurev-ecolsys-110512-135822.
95. Cancer Genome Atlas Research Network JN, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nature genetics 2013 oct;45(10):1113–20. http://www.ncbi.nlm.nih.gov/pubmed/24071849.
96. Lee JS, Das A, Jerby-Arnon L, Arafeh R, Auslander N, Davidson M, et al. Harnessing synthetic lethality to predict the response to cancer treatment. Nature communications 2018 jun;9(1):2546. http://www.ncbi.nlm.nih.gov/pubmed/29959327.

97. Kirkness EF, Haas BJ, Sun W, Braig HR, Perotti MA, Clark JM, et al. Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. Proceedings of the National Academy of Sciences 2010 jul;107(27):12168–12173. http://www.ncbi.nlm.nih.gov/pubmed/20566863.
98. Consortium TIAG. Genome Sequence of the Pea Aphid Acyrthosiphon pisum. PLoS Biology 2010 feb;8(2):e1000313. https://dx.plos.org/10.1371/journal.pbio.1000313.
99. Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M, et al. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science 2014 sep;345(6201):1181–1184. http://www.ncbi.nlm.nih.gov/pubmed/25190796.
100. Chipman AD, Ferrier DEK, Brena C, Qu J, Hughes DST, Schröder R, et al. The First Myriapod Genome Sequence Reveals Conservative Arthropod Gene Content and Genome Organisation in the Centipede Strigamia maritima. PLoS Biology 2014 nov;12(11):e1002005. http://dx.plos.org/10.1371/journal.pbio.1002005.
101. Carretero-Paulet L, Chang TH, Librado P, Ibarra-Laclette E, Herrera-Estrella L, Rozas J, et al. Genome-Wide Analysis of Adaptive Molecular Evolution in the Carnivorous Plant Utricularia gibba. Genome Biology and Evolution 2015 feb;7(2):444–456. https://doi.org/10.1093/gbe/evu288.
102. Gulia-Nuss M, Nuss AB, Meyer JM, Sonenshine DE, Roe RM, Waterhouse RM, et al. Genomic insights into the Ixodes scapularis tick vector of Lyme disease. Nature Communications 2016 apr;7(1):10507. http://www.nature.com/articles/ncomms10507.
103. Kanost MR, Arrese EL, Cao X, Chen YR, Chellapilla S, Goldsmith MR, et al. Multifaceted biological insights from a draft genome sequence of the tobacco hornworm moth, Manduca sexta. Insect Biochemistry and Molecular Biology 2016 sep;76:118–147. https://doi.org/10.1016/j.ibmb.2016.07.005.
104. Fukushima K, Fang X, Alvarez-Ponce D, Cai H, Carretero-Paulet L, Chen C, et al. Genome of the pitcher plant Cephalotus reveals genetic changes associated with carnivory. Nature Ecology & Evolution 2017 mar;1(3):0059. http://www.nature.com/articles/s41559-016-0059.
105. Rendón-Anaya M, Ibarra-Laclette E, Méndez-Bravo A, Lan T, Zheng C, Carretero-Paulet L, et al. The avocado genome informs deep angiosperm phylogeny, highlights introgressive hybridization, and reveals pathogen-influenced gene space adaptation. Proceedings of the National Academy of Sciences of the United States of America 2019 aug;116(34):17081–17089. http://www.ncbi.nlm.nih.gov/pubmed/31387975.
106. Ekblom R, Galindo J. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity 2011 jul;107(1):1–15. http://www.nature.com/articles/hdy2010152.
107. Thomson RC, Wang IJ, Johnson JR. Genome-enabled development of DNA markers for ecology, evolution and conservation. Molecular Ecology 2010 apr;19(11):2184–2195. http://www.ncbi.nlm.nih.gov/pubmed/20465588.
108. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next-generation sequencing. Nature Methods 2010 feb;7(2):111–118. http://www.ncbi.nlm.nih.gov/pubmed/20111037.

109. Turner EH, Ng SB, Nickerson DA, Shendure J. Methods for Genomic Partitioning. Annual Review of Genomics and Human Genetics 2009 sep;10(1):263–284. http://www.ncbi.nlm.nih.gov/pubmed/19630561.
110. Bybee SM, Bracken-Grissom H, Haynes BD, Hermansen RA, Byers RL, Clement MJ, et al. Targeted amplicon sequencing (TAS): a scalable next-gen approach to multilocus, multitaxa phylogenetics. Genome biology and evolution 2011;3:1312–23. http://www.ncbi.nlm.nih.gov/pubmed/22002916.
111. Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, et al. An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 2000 sep;407(6803):513–6. http://www.ncbi.nlm.nih.gov/pubmed/11029002.
112. Miller MR, Atwood TS, Eames BF, Eberhart JK, Yan YL, Postlethwait JH, et al. RAD marker microarrays enable rapid mapping of zebrafish mutations. Genome Biology 2007 jun;8(6):R105. https://doi.org/10.1186/gb-2007-8-6-r105.
113. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PLoS ONE 2008 oct;3(10):e3376. https://doi.org/10.1371/journal.pone.0003376.
114. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature Biotechnology 2009 feb;27(2):182–189. https://doi.org/10.1038/nbt.1523.
115. Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Systematic biology 2012 oct;61(5):717–26. http://www.ncbi.nlm.nih.gov/pubmed/22232343.
116. Pamilo P, Nei M. Relationships between gene trees and species trees. Mol Biol Evol 1988 sep;5(5):568–583. https://doi.org/10.1093/oxfordjournals.molbev.a040517.
117. Hurst GDD, Jiggins FM. Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. Proceedings Biological sciences / The Royal Society 2005 aug;272(1572):1525–34. https://doi.org/10.1098/rspb.2005.3056.
118. Rokas A, Williams BL, King N, Carroll SB. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 2003 oct;425(6960):798–804. http://www.ncbi.nlm.nih.gov/pubmed/14574403.
119. Brito PH, Edwards SV. Multilocus phylogeography and phylogenetics using sequence-based markers. Genetica 2009 apr;135(3):439–55. http://www.ncbi.nlm.nih.gov/pubmed/18651229.
120. Fredslund J, Madsen LH, Hougaard BK, Nielsen AM, Bertioli D, Sandal N, et al. A general pipeline for the development of anchor markers for comparative genomics in plants. BMC genomics 2006 jan;7:207. https://doi.org/10.1186/1471-2164-7-207.
121. Li C, Ortí G, Zhang G, Lu G. A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. BMC evolutionary biology 2007 jan;7(1):44. http://www.biomedcentral.com/1471-2148/7/44.

122. Lei R, Rowley TW, Zhu L, Bailey CA, Engberg SE, Wood ML, et al. PhyloMarker: A Tool for Mining Phylogenetic Markers Through Genome Comparison: Application of the Mouse Lemur (Genus Microcebus) Phylogeny. Evolutionary Bioinformatics 2012 jul;p. 423. https://doi.org/10.4137/EBO.S9886.
123. Davey JW, Cezard T, Fuentes-Utrilla P, Eland C, Gharbi K, Blaxter ML. Special features of RAD Sequencing data: implications for genotyping. Molecular Ecology 2013 jun;22(11):3151–3164. http://doi.wiley.com/10.1111/mec.12084.
124. Arnold B, Corbett-Detig RB, Hartl D, Bomblies K. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Molecular Ecology 2013 jun;22(11):3179–3190. http://doi.wiley.com/10.1111/mec.12276.
125. Karl Sa, Avise JC. PCR-based assays of mendelian polymorphisms from anonymous single-copy nuclear DNA: techniques and applications for population genetics. Molecular biology and evolution 1993 mar;10(2):342–61. http://www.ncbi.nlm.nih.gov/pubmed/8098128.
126. Bertozzi T, Sanders KL, Sistrom MJ, Gardner MG. Anonymous nuclear loci in non-model organisms : making the most of high-throughput genome surveys. Bioinformatics 2012;28(14):1807–10. http://www.ncbi.nlm.nih.gov/pubmed/22581180.
127. Green P. 2X Genomes–Does Depth Matter? Genome research 2007 nov;17(11):1547–9. http://www.ncbi.nlm.nih.gov/pubmed/17975171.
128. World Spider Catalog, World Spider Catalog (2018).; 2018. http://wsc.nmbe.ch.
129. Cooke J. Spider Genus Dysdera (Araneae, Dysderidae). Nature 1965 mar;205(4975):1027–1028. http://www.nature.com/doifinder/10.1038/2051027b0.
130. Cooke J. A Contribution to the Biology of the British Spiders belonging to the Genus Dysdera. Oikos 1965;16(1/2):20. http://www.jstor.org/stable/3564861?origin=crossref.
131. Deeleman-Reinhold CL. The genus Rhode and the harpacteine genera Stalagtia, Folkia, Minotauria, and Kaemis (Araneae, Dysderidae) of Yugoslavia and Crete, with remarks on the genus Harpactea. Revue Arachnologique 1993;10(6):105–135.
132. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Molecular Biology and Evolution 2017 jul;34(7):1812–1819. https://doi.org/10.1093/molbev/msx116.
133. Carlson DE, Hedin M. Comparative transcriptomics of Entelegyne spiders (Araneae, Entelegynae), with emphasis on molecular evolution of orphan genes. PLOS ONE 2017 apr;12(4):e0174102. https://doi.org/10.1371/journal.pone.0174102.
134. Ballesteros JA, Sharma PP. A Critical Appraisal of the Placement of Xiphosura (Chelicerata) with Account of Known Sources of Phylogenetic Error. Systematic Biology 2019 mar;https://doi.org/10.1093/sysbio/syz011.
135. Lozano-Fernandez J, Tanner AR, Giacomelli M, Carton R, Vinther J, Edgecombe GD, et al. Increasing species sampling in chelicerate genomic-scale datasets provides support for monophyly of Acari and Arachnida. Nature Communications 2019;10:2295. https://doi.org/10.1038/s41467-019-10244-7.

136. Fernández R, Kallal RJ, Dimitrov D, Ballesteros JA, Arnedo MA, Giribet G, et al. Phylogenomics, Diversification Dynamics, and Comparative Transcriptomics across the Spider Tree of Life. Current Biology 2018 may;28(9):1489–1497.e5. https://doi.org/10.1016/j.cub.2018.03.064.
137. Anguita F, Hernán F. The Canary Islands origin: a unifying model. Journal of Volcanology and Geothermal Research 2000 dec;103(1-4):1–26. https://doi.org/10.1016/S0377-0273(00)00195-5.
138. Garb JE, Sharma PP, Ayoub NA. Recent progress and prospects for advancing arachnid genomics. Current Opinion in Insect Science 2018 feb;25:51–57. https://doi.org/10.1016/j.cois.2017.11.005.
139. Sanggaard KW, Bechsgaard JS, Fang X, Duan J, Dyrlund TF, Gupta V, et al. Spider genomes provide insight into composition and evolution of venom and silk. Nature communications 2014 may;5(May):3765. https://doi.org/10.1038/ncomms4765.
140. Schwager EE, Sharma PP, Clarke T, Leite DJ, Wierschin T, Pechmann M, et al. The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. BMC Biology 2017 dec;15(1):62. https://doi.org/10.1186/s12915-017-0399-x.
141. King GF, Hardy MC. Spider-Venom Peptides: Structure, Pharmacology, and Potential for Control of Insect Pests. Annual Review of Entomology 2013 jan;58(1):475–496. https://doi.org/10.1146/annurev-ento-120811-153650.
142. Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, et al. The amphioxus genome and the evolution of the chordate karyotype. Nature 2008 jun;453(7198):1064–1071. https://doi.org/10.1038/nature06967.
143. Kenny NJ, Chan KW, Nong W, Qu Z, Maeso I, Yip HY, et al. Ancestral whole-genome duplication in the marine chelicerate horseshoe crabs. Heredity 2016 feb;116(2):190–199. http://www.nature.com/articles/hdy201589.
144. Lozano-Fernandez J, Carton R, Tanner AR, Puttick MN, Blaxter M, Vinther J, et al. A molecular palaeobiological exploration of arthropod terrestrialization. Philosophical transactions of the Royal Society of London Series B, Biological sciences 2016 jul;371(1699):20150133. http://www.ncbi.nlm.nih.gov/pubmed/27325830.
145. Vizueta J, Frias-López C, Macías-Hernández N, Arnedo MA, Sánchez-Gracia A, Rozas J, et al. Evolution of chemosensory gene families in arthropods: Insight from the first inclusive comparative transcriptome analysis across spider appendages. Genome Biology and Evolution 2017 dec;9(1):178–196. https://doi.org/10.1093/gbe/evw296.
146. Shaw KL, Gillespie RG. Comparative phylogeography of oceanic archipelagos: Hotspots for inferences of evolutionary process. Proceedings of the National Academy of Sciences of the United States of America 2016 jul;113(29):7986–93. http://www.ncbi.nlm.nih.gov/pubmed/27432948.
147. Whittaker RJ, Fernández-Palacios JM, Matthews TJ, Borregaard MK, Triantis KA. Island biogeography: Taking the long view of nature's laboratories. Science 2017 sep;357(6354):eaam8326. https://doi.org/10.1126/science.aam8326.

148. Gómez A, González-Martínez SC, Collada C, Climent J, Gil L. Complex population genetic structure in the endemic Canary Island pine revealed using chloroplast microsatellite markers. TAG Theoretical and Applied Genetics 2003 oct;107(6):1123–1131. https://doi.org/10.1007/s00122-003-1320-2.
149. Kim SC, Crawford DJ, Francisco-Ortega J, Santos-Guerra A. Plant Systematics and Evolution Adaptive radiation and genetic differentiation in the woody Sonchus alliance (Asteraceae: Sonchinae) in the Canary Islands. Pl Syst Evol 1999;215:101–118. https://www.jstor.org/stable/23643354.
150. González P, Pinto F, Nogales M, Jiménez-asensio J, Hernández M, Cabrera VM. Phylogenetic Relationships of the Canary Islands Endemic Lizard GenusGallotia(Sauria: Lacertidae), Inferred from Mitochondrial DNA Sequences. Molecular Phylogenetics and Evolution 1996 aug;6(1):63–71. http://www.ncbi.nlm.nih.gov/pubmed/8812306.
151. Brown RP, Pestano J. Phylogeography of skinks (Chalcides) in the Canary Islands inferred from mitochondrial DNA sequences. Molecular ecology 1998 sep;7(9):1183–91. http://www.ncbi.nlm.nih.gov/pubmed/9734075.
152. Juan C, Emerson BC, Oromi P, Hewitt GM. Colonization and diversification: towards a phylogeographic synthesis for the Canary Islands. Trends in Ecology & Evolution 2000 mar;15(3):104–109. http://www.cell.com/article/S0169534799017760/fulltext.
153. Avanzati AM, Baratti M, Bernini F. Molecular and morphological differentiation between steganacarid mites (Acari: Oribatida) from the Canary islands. Biological Journal of the Linnean Society 1994 aug;52(4):325–340. https://doi.org/10.1111/j.1095-8312.1994.tb00995.x.
154. Arnedo MA, Ribera C. Radiation of the genus Dysdera (Araneae, Haplogynae, Dysderidae) in the Canary Islands: The island of Gran Canaria. Zoologica Scripta 1997 jul;26(3):205–243. http://doi.wiley.com/10.1111/j.1463-6409.1997.tb00413.x.
155. Arnedo MA, Oromi P, Ribera C. Radiation of the Spider Genus Dysdera (Araneae, Dysderidae) in the Canary Islands: Cladistic Assessment Based on Multiple Data Sets. Cladistics 2001;17:313–353. http://doi.wiley.com/10.1006/clad.2001.0168.
156. Bidegaray-Batista L, Macias-Hernandez N, Oromi P, Arnedo MA. Living on the edge: Demographic and phylogeographical patterns in the woodlouse-hunter spider Dysdera lancerotensis Simon, 1907 on the eastern volcanic ridge of the Canary Islands. Molecular Ecology 2007 aug;16(15):3198–3214. https://doi.org/10.1111/j.1365-294X.2007.03351.x.
157. Macias-Hernandez N, de la Cruz Lopez S, Roca-Cusachs M, Oromi P, Arnedo MA. A geographical distribution database of the genus Dysdera in the Canary Islands (Araneae, Dysderidae). ZooKeys 2016 oct;625(625):11–23. https://doi.org/10.3897/zookeys.625.9847.
158. Macias-Hernandez N, Oromi P, Arnedo MA. Integrative taxonomy uncovers hidden species diversity in woodlouse hunter spiders (Araneae, Dysderidae) endemic to the Macaronesian archipelagos. Systematics and Biodiversity 2010 dec;8(4):531–553. https://doi.org/10.1080/14772000.2010.535865.
159. Macias-Hernandez N, Bidegaray-Batista L, Emerson BC, Oromi P, Arnedo MA. The imprint of geologic history on within-island diversification of woodlouse-hunter spiders (Araneae, Dysderidae) in the canary islands. Journal of Heredity 2013;104(3):341–356.

160. Řezáč M, Pekár S. Evidence for woodlice-specialization in Dysdera spiders: Behavioural versus developmental approaches. Physiological Entomology 2007;32(4):367–371. https://doi.org/10.1111/j.1365-3032.2007.00588.x.
161. Řezác M, Pekár S, Lubin Y. How oniscophagous spiders overcome woodlouse armour. Journal of Zoology 2008;275(1):64–71. https://doi.org/10.1111/j.1469-7998.2007.00408.x.
162. Wieser W, Dallinger R, Busch G. The flow of copper through a terrestrial food chain. Oecologia 1977 sep;30(3):265–272. https://doi.org/10.1007/BF01833633.
163. Paoletti MG, Hassall M. Woodlice (Isopoda: Oniscidea): their potential for assessing sustainability and use as bioindicators. Agriculture, Ecosystems & Environment 1999 jun;74(1-3):157–165. https://doi.org/10.1016/S0167-8809(99)00035-3.
164. Hopkin SP, Martin MH. Assimilation of Zinc, Cadmium, Lead, Copper, and Iron by the Spider Dysdera crocata, a Predator of Woodlice. Bull Environ Contam Toxicol 1985;34:183–187.
165. Pekár S, Toft S. Trophic specialisation in a predatory group: the case of prey-specialised spiders (Araneae). Biological Reviews 2015 aug;90(3):744–761. https://doi.org/10.1111/brv.12133.
166. Pekár S, Líznarová E, Řezác M. Suitability of woodlice prey for generalist and specialist spider predators: a comparative study. Ecological Entomology 2016 apr;41(2):123–130. http://doi.wiley.com/10.1111/een.12285.
167. Toft S, Macías-Hernández N. Metabolic adaptations for isopod specialization in three species of Dysdera spiders from the Canary Islands. Physiological Entomology 2017 jun;42(2):191–198. http://doi.wiley.com/10.1111/phen.12192.
168. Sharma PP, Kaluziak ST, Pérez-Porro AR, González VL, Hormiga G, Wheeler WC, et al. Phylogenomic interrogation of arachnida reveals systemic conflicts in phylogenetic signal. Molecular biology and evolution 2014 nov;31(11):2963–84. http://www.ncbi.nlm.nih.gov/pubmed/25107551.
169. Kallal RJ, Fernández R, Giribet G, Hormiga G. A phylotranscriptomic backbone of the orb-weaving spider family Araneidae (Arachnida, Araneae) supported by multiple methodological approaches. Molecular Phylogenetics and Evolution 2018 sep;126:129–140. https://doi.org/10.1016/j.ympev.2018.04.007.
170. Andersson M, Jia Q, Abella A, Lee XY, Landreh M, Purhonen P, et al. Biomimetic spinning of artificial spider silk from a chimeric minispidroin. Nature Chemical Biology 2017 mar;13(3):262–264. http://www.nature.com/articles/nchembio.2269.
171. Frıas-Lopez C, Sanchez-Herrero JF, Guirao-Rico S, Mora E, Arnedo MA, Sanchez-Gracia A, et al. Domino: Development of informative molecular markers for phylogenetic and genome-wide population genetic studies in non-model organisms. Bioinformatics 2016 aug;32(24):3753–3759. https://doi.org/10.1093/bioinformatics/btw534.
172. Thomson RC, Shedlock AM, Edwards SV, Shaffer HB. Developing markers for multilocus phylogenetics in non-model organisms: A test case with turtles. Molecular phylogenetics and evolution 2008 nov;49(2):514–25. http://www.ncbi.nlm.nih.gov/pubmed/18761096.

173. i5k Consortium. The i5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment. Journal of Heredity 2013 sep;104(5):595–600. https://doi.org/10.1093/jhered/est050.
174. Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al. Earth BioGenome Project: Sequencing life for the future of life. Proceedings of the National Academy of Sciences of the United States of America 2018 apr;115(17):4325–4333. http://www.ncbi.nlm.nih.gov/pubmed/29686065.
175. Sánchez-Herrero JF, Frías-López C, Escuer P, Hinojosa-Alvarez S, Arnedo MA, Sánchez-Gracia A, et al. The draft genome sequence of the spider Dysdera silvatica (Araneae, Dysderidae): A valuable resource for functional and evolutionary genomic studies in chelicerates. GigaScience 2019 aug;8(8). https://doi.org/10.1093/gigascience/giz099.
176. Babb PL, Lahens NF, Correa-Garhwal SM, Nicholson DN, Kim EJ, Hogenesch JB, et al. The Nephila clavipes genome highlights the diversity of spider silk genes and their complex expression. Nature Genetics 2017 jun;49(6):895–903. http://www.nature.com/articles/ng.3852.
177. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;31(19):3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

Anexo

Apéndice A

Articulos resultados de colaboraciones científicas

Este apartado incluye otras publicaciones en las que el doctorando ha participado pero no ha incluido como parte de esta tesis doctoral.

A

A.1. *Streptococcus gallolyticus subsp. gallolyticus* from human and animal origins: genetic diversity, antimicrobial susceptibility, and characterization of a vancomycin-resistant calf isolate carrying a vanA-Tn1546-like element

El objetivo de este estudio fue la caracterización de la susceptibilidad a antibióticos y diversidad genética de 41 aislados de *Streptococcus gallolyticus subsp. gallolyticus*, 18 de los cuales fueron obtenidos de animales y 23 de ellos de muestras clínicas. La susceptibilidad a antibióticos fue determinada por un sistema semi-automático Wider y la diversidad genética por electroforesis de campo pulsado (PFGE) con *SmaI*. Los aislados de animales se agrupan por separado en el análisis PFGE pero no se encuentran diferencias estadísticamente significativas para el análisis de resistencias a antibióticos entre los dos grupos.

La cepa LMG 17956 tipificada como 28 (ST28), recuperada de excrementos de un ternero, presentaba unos niveles elevados de resistencia a vancomicina y teicoplanina (MIC >256 mg/l). Su mecanismo de resistencia a glicopéptidos, caracterizado por hibridación Southern Blot y una estrategia de *"primer walking"*, así como finalmente el genoma, mediante secuenciación masiva, fue comparado con otros 4 genomas relacionados de *S. gallolyticus subsp. gallolyticus*. Experimentos de hibridación mostraron como un elemento similar a Tn1546 estaba integrado en el cromosoma bacteriano. En corcondancia con estos resultados, la secuenciación masiva confirmó una delección parcial de la región *vanY-vanZ* y una duplicación parcial del gen *vanH*. El análisis genómico comparativo reveló que la cepa LMG 17956 ST28 habría adquirido un inusual número de elementos transponibles y habría experimentando importantes reordenamientos cromosómicos así como ganancia y pérdida de genes.

En conclusión, los aislados de *S. gallolyticus subsp. gallolyticus* de origen animal parecen tener linajes separados de aquellos que infectan humanos. Además, presentamos un aislado resistente a glicopéptidos de origen vacuno que posee un elemento Tn1546-like integrado en su cromosoma.

A

AAC Journals.ASM.org

Streptococcus gallolyticus subsp. *gallolyticus* from Human and Animal Origins: Genetic Diversity, Antimicrobial Susceptibility, and Characterization of a Vancomycin-Resistant Calf Isolate Carrying a *vanA*-Tn*1546*-Like Element

Beatriz Romero-Hernández,[a] Ana P. Tedim,[a,b] José Francisco Sánchez-Herrero,[c] Pablo Librado,[c] Julio Rozas,[c] Gloria Muñoz,[d] Fernando Baquero,[a,b] Rafael Cantón,[a,e] Rosa del Campo,[a,e] the Spanish Network for Research on Infectious Diseases (REIPI)

Servicio de Microbiología, Hospital Universitario Ramón y Cajal, and IRYCIS, Madrid, Spain[a]; CIBER en Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain[b]; Departament de Genètica i Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain[c]; UCAGT, IRYCIS, Madrid, Spain[d]; Red Española de Investigación en Patología Infecciosa (REIPI), Seville, Spain[e]

The aim of this work was to characterize the antibiotic susceptibility and genetic diversity of 41 *Streptococcus gallolyticus* subsp. *gallolyticus* isolates: 18 isolates obtained from animals and 23 human clinical isolates. Antibiotic susceptibility was determined by the semiautomatic Wider system and genetic diversity by pulsed-field gel electrophoresis (PFGE) with SmaI. Animal isolates grouped separately in the PFGE analysis, but no statistical differences in antimicrobial resistance were found between the two groups. The LMG 17956 sequence type 28 (ST28) strain recovered from the feces of a calf exhibited high levels of resistance to vancomycin and teicoplanin (MIC, ≥256 mg/liter). Its glycopeptide resistance mechanism was characterized by Southern blot hybridization and a primer-walking strategy, and finally its genome, determined by whole-genome sequencing, was compared with four closely related *S. gallolyticus* subsp. *gallolyticus* genomes. Hybridization experiments demonstrated that a Tn*1546*-like element was integrated into the bacterial chromosome. In agreement with this finding, whole-genome sequencing confirmed a partial deletion of the *vanY-vanZ* region and partial duplication of the *vanH* gene. The comparative genomic analyses revealed that the LMG 17956 ST28 strain had acquired an unusually high number of transposable elements and had experienced extensive chromosomal rearrangements, as well as gene gain and loss events. In conclusion, *S. gallolyticus* subsp. *gallolyticus* isolates from animals seem to belong to lineages separate from those infecting humans. In addition, we report a glycopeptide-resistant isolate from a calf carrying a Tn*1546*-like element integrated into its chromosome.

New taxonomic criteria have recently been applied to the *Streptococcus bovis/equinus* complex, mainly on the basis of the genetic diversity of the *sodA* gene, which is considered the best target for adequate identification (1). These taxonomic advances have permitted the study of the epidemiological correlations between particular subspecies and specific human pathologies (2), including gastrointestinal colonization by *Streptococcus gallolyticus* subsp. *gallolyticus* (formerly *Streptococcus bovis* biotype I) and its coincidence with colorectal cancer (3).

The rates of colonization by *S. gallolyticus* subsp. *gallolyticus* are around 5 to 10% in humans but could be much higher in animals (4). Bacteremia and endocarditis are the main relevant clinical manifestations in humans and in avian species (5, 6). The mortality rate in a broiler flock outbreak was 4.3% (7), and the organism has also been implicated in other veterinary pathologies (8, 9).

Similarities and differences between *S. gallolyticus* subsp. *gallolyticus* isolates from humans and animals have been described previously (8, 10), including the existence of particular virulent clones with increased invasion and adherence abilities, which favor bloodstream infections (11–13). A recent multilocus sequence typing (MLST)-based study demonstrated a lack of specificity for any particular host or geographical location (10).

Antibiotic resistance is a not infrequent feature of *S. gallolyticus* subsp. *gallolyticus* isolates. Resistance to macrolides and tetracyclines and high-level resistance to aminoglycosides are most commonly reported (14–16). Resistance to glycopeptides has been reported occasionally as a consequence of the acquisition of enterococcal *vanA* or *vanB* mechanisms (17–19).

Transmission of *S. gallolyticus* subsp. *gallolyticus* strains between animals and humans has not been documented yet, although in view of the example of the related genus *Enterococcus*, zoonotic potential cannot be ruled out. A major prevalence of endocarditis caused by *S. bovis* in rural areas of France and Spain has been demonstrated in this sense (20, 21).

We have characterized an *S. gallolyticus* subsp. *gallolyticus* collection including isolates recovered from animals and humans and have compared their genetic diversity and antimicrobial susceptibility profiles. Furthermore, we have characterized a glyco-

Received 13 August 2014 **Returned for modification** 20 November 2014
Accepted 13 January 2015

Accepted manuscript posted online 20 January 2015

Citation Romero-Hernández B, Tedim AP, Sánchez-Herrero JF, Librado P, Rozas J, Muñoz G, Baquero F, Cantón R, del Campo R, the Spanish Network for Research on Infectious Diseases (REIPI). 2015. *Streptococcus gallolyticus* subsp. *gallolyticus* from human and animal origins: genetic diversity, antimicrobial susceptibility, and characterization of a vancomycin-resistant calf isolate carrying a *vanA*-Tn*1546*-like element. Antimicrob Agents Chemother 59:2006–2015. doi:10.1128/AAC.04083-14.

Address correspondence to Rosa del Campo, rosacampo@yahoo.com.

doi:10.1128/AAC.04083-14

TABLE 1 Main characteristics of the 18 animal isolates

Isolate	Origin	Yr	Country	ST[a]	Antibiotic resistance	
					Phenotype[b]	Genotype
LMG 14621	Dove abscess	1994	Belgium	19	Fo, Lv	
LMG 14622	Dove	1994	Belgium	19	Fo	
LMG 14623	Dove liver	1994	Belgium	20	Fo, Er, Cd, Mn, St	*erm*(B), *tet*(M)
LMG 14634	Cow	1994	Belgium	7	Fo, Er, Cd, Mn, St	*erm*(B), *tet*(M)
LMG 14821	Dove gut	1994	Belgium	21	Fo, Cd, Mn, St	*tet*(M)
LMG 14823	Dove gut	1994	Belgium	22	Fo, Er, Cd, Mn, St	*erm*(B), *tet*(M)
LMG 14855	Horse gut	1994	Belgium	23		
LMG 14856	Horse gut	1994	Belgium	24		
LMG 14870	Cow gut	1994	Belgium	25	Er, Cd, St	*erm*(B)
LMG 14876	Cow tonsil	1994	Belgium	26	Fo, Er, Cd, Mn, St	*erm*(B), *tet*(M)
LMG 14878	Pig lung	1994	Belgium	19	Fo, Er	
LMG 15572	Goat rumen	1994	Australia	27		
LMG 15573	Goat rumen	1994	Australia	27		
LMG 16005	Cow gut	1995	Belgium	3	Er, Cd, Mn, SxT, St	*erm*(B), *tet*(M)
LMG 17956	Cow gut	1997	Netherlands	28	Cd, Q/D, SxT, Va	Tn*1546*-like
LMG 22782	Dog	2000	Belgium	17	Er, Cd, Mn, Q/D	*erm*(B), *tet*(M)
05WDK43740 002	Cow gut	Unknown	Unknown	13	Fo	
DSM16831	Koala feces	1990	Australia	1	Fo, Gm	*aac*(6′)-*aph*(2″)

[a] ST, sequence type.
[b] Fo, fosfomycin; Lv, levofloxacin; Er, erythromycin; Cd, clindamycin; Mn, minocycline; St, streptomycin at 1,000 mg/liter; SxT, trimethoprim-sulfamethoxazole; Q/D, quinupristin-dalfopristin; Va, vancomycin; Gm, gentamicin at 500 mg/liter.

peptide-resistant isolate carrying a *vanA*-Tn*1546*-like element that was recovered from the feces of a calf.

MATERIALS AND METHODS

Bacterial isolates. A total of 41 *S. gallolyticus* subsp. *gallolyticus* isolates (18 from animals and 23 from humans) were included in the study (Tables 1 and 2; see Fig. 1). Some of the human and animal isolates were kindly provided from the public BCCM/LMG strain collection (http://bccm.belspo.be/about/lmg.php) (13), whereas the bacteremic isolates were obtained at the Ramón y Cajal University Hospital (1) in Madrid, Spain. All strains were identified to the subspecies level by PCR amplification of an internal fragment of the *sodA* gene and further nucleotide

TABLE 2 Main characteristics of the 23 human-invasive isolates

Isolate	Origin	Yr	Country	ST[a]	Antibiotic resistance	
					Phenotype[b]	Genotype
003080/00	Human gut	2000	Germany	5	Fo, Gm	*aac*(6′)-*aph*(2″)
005950/03	Human heart valve	2003	Germany	11	Fo, Er, Cd, Mn, St, Q/D	*erm*(B), *tet*(M)
006718/00		2000	Germany	7	Mn	*tet*(M)
007849/02		2002	Germany	8	Fo, Er, Cd, Lv, St, Q/D	*erm*(B)
010288/01		2001	Germany	3	Er, Cd, Mn, St	*erm*(B), *tet*(M)
010672/01		2001	Germany	6	Mn, Gm	*tet*(M), *aac*(6′)-*aph*(2″)
021702/06		2006	Germany	9	Fo, Mn, Gm	*tet*(M), *aac*(6′)-*aph*(2″)
12932/01		2001	Germany	3	Fo, Er, Cd, Mn	*erm*(B), *tet*(M)
B1	Bacteremia	2004	Spain	34	SxT	
B6		2009	Spain	35	Fo, Cd, St	
B11		2005	Spain	7	Fo, Er, Cd, St, Lv, Q/D	*erm*(B)
B13		2004	Spain	36	Fo, Er, Cd, St, Lv	*erm*(B)
B14		2004	Spain	5		
B19		2006	Spain	7	Fo, Er, Cd, Lv, St, Q/D	*erm*(B)
B22		2006	Spain	37		
B28		2009	Spain	38	Fo, Mn, Gm	*erm*(B), *aac*(6′)-*aph*(2″)
B29		2009	Spain	39	Er, Cd, Mn	*erm*(B), *tet*(M)
B35		2003	Spain	26	Er, Cd, St	
B41		2010	Spain	5	Er	*erm*(B)
B49		2003	Spain	6		
B51		2003	Spain	40	Fo	
B52		2003	Spain	5	Er, Cd	*erm*(B)
K6236/35_MS		Unknown	Unknown	12		

[a] ST, sequence type.
[b] Fo, fosfomycin; Gm, gentamicin at 500 mg/liter; Er, erythromycin; Cd, clindamycin; Mn, minocycline; St, streptomycin at 1,000 mg/liter; Q/D, quinupristin-dalfopristin; Lv, levofloxacin; SxT, trimethoprim sulfamethoxazole.

sequencing (1). Matrix-assisted laser desorption ionization–time of flight (MALDI-TOF) mass spectrometry (MS) using the Bruker Biotyper system (Bruker Daltonics, Bremen, Germany) was also performed as part of the bacterial identification scheme (22).

Antimicrobial susceptibility testing. Susceptibility was determined using the Wider semiautomatic microdilution system (Fco. Soria Melguizo, Madrid, Spain), and results were interpreted according to the guidelines of the Clinical and Laboratory Standards Institute by using the criteria described for enterococci (23). The presence or absence of the *erm*(B), *tet*(M), and *aac*(6′)-*aph*(2″) genes was determined by PCR using specific primers [*erm*(B)-F (GAAAAGTACTCAACCAAATA) and *erm*(B)-R (AGTAACGGTACTTAAATTGTTTA), *tet*(M)-F (GTTAAATAGTGTTCTTGGAG) and *tet*(M)-R (CTAAGATATGGCTCTAACAA), and *aac-aph*-F (CCAAGAGCAATAAGGCATA) and *aac-aph*-R (CACTATCATAACCACTACCG)].

Genetic diversity. Clonal relatedness was determined by pulsed-field gel electrophoresis (PFGE) with SmaI by using a protocol initially described for *Streptococcus suis* serotype 2 (24). We constructed a dendrogram using Phoretix software, version 5.0 (Nonlinear Dynamics Ltd., Newcastle, United Kingdom), based on the Dice coefficient.

Tn*1546* characterization. A primer-walking scheme described previously was used to characterize Tn*1546* (25). The location of the glycopeptide resistance mechanism was assessed by hybridization of the I-CeuI-digested genomic DNA of the LMG 17956 ST28 strain, and of positive- and negative-control strains (the *vanA*-containing strains *Enterococcus faecalis* RC715 and *Enterococcus faecium* RC714 and the vancomycin-susceptible strain *S. gallolyticus* subsp. *gallolyticus* ATCC BA-2069, respectively), with probes for the 16S rRNA genes and *vanA*.

Initially, the nitrocellulose membrane was hybridized with the 16S rRNA gene probes obtained from the *E. faecalis* and *E. faecium* control strains and the glycopeptide-resistant *S. gallolyticus* isolate after universal PCR with V3–V4 primers.

In a second stage, the same membrane, after adequate washes, was newly hybridized with a *vanA* probe from a PCR *vanA* fragment obtained from an *Enterococcus faecium* control strain (25). All probes were labeled by random primer labeling with Redivue [^{32}P]dCTP (Amersham, Little Chalfont, Buckinghamshire, United Kingdom). Prehybridization and hybridization were carried out in Rapid buffer (Amersham) at 60°C for 30 min and at 56°C for 18 h, respectively. The membrane was washed twice at 56°C in 2× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate)–0.1% SDS and then twice at room temperature with 1× SSC–0.1% SDS and 0.7× SSC–0.1% SDS, successively. Autoradiography was carried out by filter exposure for 72 h at −80°C.

Conjugative transfer of vancomycin resistance and plasmid detection. We tested the ability of vancomycin resistance to be transferred by conjugation by using the *E. faecalis* JH2-2 and *E. faecium* OG-RF1 strains (both of which are resistant to rifampin and fusidic acid) as recipients. Conjugation was developed by the filter mating method, and transconjugants were selected onto m-*Enterococcus* agar supplemented with 25 mg/liter of fusidic acid, 30 mg/liter of rifampin, and 16 mg/liter of vancomycin. The enterococcal model was chosen for conjugation because enterococci are the natural hosts for the vancomycin resistance determinants and also because of the lack of a validated *S. gallolyticus* subsp. *gallolyticus* recipient strain.

The presence of plasmids in the donor and recipient strains was explored after a plasmid extraction protocol and subsequent electrophoresis.

Whole-genome sequencing. Total DNA from the *vanA*-containing *S. gallolyticus* subsp. *gallolyticus* LMG 17956 ST28 isolate was obtained by using a QIAamp kit (Qiagen, Hilden, Germany). A library was generated with 100 ng of DNA by using the Xpress Plus Fragment Library kit (Life Technologies, Eggenstein-Leopoldshafen, Germany). Quality was measured by the High Sensitivity DNA kit in the 2200 TapeStation system (Agilent Technologies, Palo Alto, CA, USA). Pyrosequencing was performed by using Ion Torrent technology with an Ion 316 v2 chip. The preliminary assembly obtained with MIRA software (26) was further completed by Era7 Bioinformatics. Protein-coding genes were annotated by performing a BLAST search against the COGs (clusters of orthologous groups) database (http://www.ncbi.nlm.nih.gov/COG/) (27, 28).

Comparison of genomes. The genome sequence of the *S. gallolyticus* subsp. *gallolyticus* LMG 17956 ST28 isolate was compared with the previously sequenced genomes of *S. gallolyticus* subsp. *gallolyticus* UCN34 (BioProject accession no. PRJNA46061), *S. gallolyticus* subsp. *gallolyticus* ATCC 43143 (PRJDA162103), and *S. gallolyticus* subsp. *gallolyticus* ATCC BAA-2069 (PRJNA63617). Chromosomal rearrangements were inferred by aligning the genome sequences of these strains with MAUVE software (29). Groups of orthologous genes were defined by the OrthoMCL algorithm (30). Species-specific (exclusive) orthologous gene groups were represented using the VennDiagram package (version 1.6.5) of the R programming language (31, 32).

Gene family expansions. Orthologous groups that had experienced significant gene expansion in the LMG 17956 lineage were identified using the probabilistic and phylogenetic framework provided by BadiRate software (33). In addition to the four *S. gallolyticus* strains, two closely related outgroup species were included in this analysis: *Streptococcus equinus* and *Streptococcus thermophilus*. In particular, the multiple-sequence alignments (MSAs) of their 1:1 orthologous protein-coding genes (a single gene copy per strain/species) were built with MAFFT, version 7 (34). These MSAs were concatenated using in-house-developed Perl scripts, leading to a concatenation that comprised 996,954 positions. The phylogenetic relationships among the six strains/species were estimated from this concatenation by RAxML, version 7.2.8, using a general time-reversible (GTR) model of DNA substitution with gamma distribution, with *S. thermophilus* as the outgroup species (to root the phylogenomic tree) (35, 36).

Statistical analysis. The chi-square test was used for the antibiotic resistance analysis.

Nucleotide sequence accession numbers. The genome of the vancomycin-resistant *S. gallolyticus* subsp. *gallolyticus* LMG 17956 ST28 isolate was deposited in the European Nucleotide Archive (ENA) under accession numbers CCBC010000001 to CCBC010000260 (http://www.ebi.ac.uk/ena/data/view/CCBC010000101).

RESULTS

Genetic diversity analysis. The genetic diversity analysis based on the dendrogram calculated from the PFGE patterns clustered all isolates into five distinct groups (Fig. 1). Human bacteremic isolates clustered in groups II, III, and IV, whereas almost all animal isolates clustered in group I. It should be noted that the glycopeptide-resistant LMG 17956 ST28 strain, obtained from the feces of a calf, was grouped with the bacteremic human isolates of group IV (Fig. 1).

Antimicrobial susceptibility. In general, all isolates presented low resistance rates, with no significant differences between animal and human isolates (Table 3). The most relevant result was the glycopeptide resistance (MIC values of both vancomycin and teicoplanin, 256 mg/liter) detected in the *S. gallolyticus* subsp. *gallolyticus* LMG 17956 ST28 strain obtained from the feces of a calf.

Molecular characterization of glycopeptide resistance. We were unable to transfer vancomycin resistance by conjugation into the *E. faecalis* JH2-2 and *E. faecium* OG-RF1 enterococcal recipients after independent experiments. Moreover, the presence of plasmids was ruled out for both the donor and the recipient strains after independent negative plasmid extractions.

Initially, we followed the primer-walking strategy approximation to characterize the glycopeptide resistance mechanism. Using this technique, we identified the first 10,850 bp of Tn*1546*, but amplification of the *vanY-vanZ* region was not possible. Positive

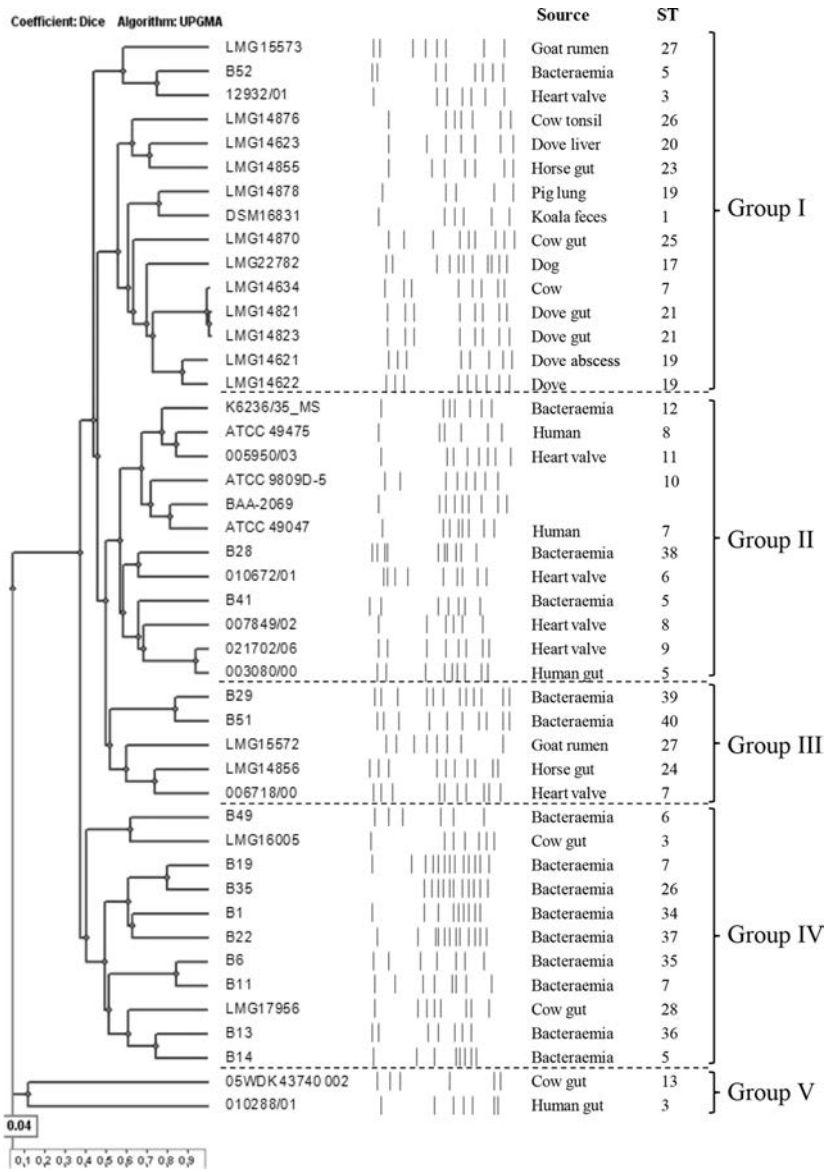


FIG 1 Dendrogram based on the Dice coefficient of the PFGE patterns. The source and sequence type (ST) of each strain are also given.

hybridization in the same I-CeuI fragment was observed with both the *vanA* and 16S rRNA genes, indicating integration of the Tn*1546*-like element into the bacterial chromosome (Fig. 2).

Whole-genome sequencing of the vancomycin-resistant *S. gallolyticus* subsp. *gallolyticus* LMG 17956 ST28 strain. For further characterization of the Tn*1546* element, we sequenced the whole genome of the vancomycin-resistant *S. gallolyticus* subsp. *gallolyticus* LMG 17956 ST28 strain. This genome was deposited in the European Nucleotide Archive.

Bioinformatic analysis of contigs in comparison with the canonical Tn*1546* sequence from the work of Arthur et al. (37) demonstrated an inversion of the proximal part of the transposon and a deletion of 1,835 bp in the distal Tn*1546*-like fragment (from bp 9016 to 10851), including the *vanY* and *vanZ* genes (Fig. 3).

The genome assembly comprises a total of 2,698,137 bp, with 2,410 genes, 58 tRNAs, 4 copies of rRNA genes, and 1 other noncoding RNA. In agreement with the analysis of clusters of orthologous groups (COGs), 35% of the genetic content was related to metabolic processes and 28% to information storage and processing, while another 17% of the genetic content contributed to cel-

TABLE 3 Comparison of antimicrobial resistance in animal and human isolates

Antibiotic[a]	No. (%) of isolates with resistance		
	Animal isolates ($n = 18$)	Human isolates ($n = 23$)	P^b
Fosfomycin	10 (55.5)	11 (47.8)	NS
Erythromycin	8 (44.4)	11 (47.8)	NS
Clindamycin	9 (50.0)	11 (47.8)	NS
Minocycline	7 (38.8)	8 (34.7)	NS
Gentamicin (500 mg/liter)	1 (5.5)	4 (17.3)	NS
Streptomycin (1,000 mg/liter)	7 (38.8)	8 (34.7)	NS
SxT	2 (11.1)	1 (4.3)	NS
Q/D	2 (11.1)	4 (17.3)	NS
Levofloxacin	1 (5.5)	4 (17.3)	NS
Vancomycin	1 (5.5)	0 (0)	NS

[a] SxT, trimethoprim-sulfamethoxazole; Q/D, quinupristin-dalfopristin.
[b] Statistical significance of differences in antimicrobial resistance between animal and human isolates. NS, not significant.

lular processes and signaling. However, 32% of the genome was poorly characterized, including those genes for which assignation to COGs was not possible (Fig. 4).

Comparison of the four *S. gallolyticus* strains revealed that the LMG 17956 ST28 genome presented extensive rearrangements, contained a high number of transposase-encoding genes ($n = 111$) (mainly related to IS*1167*, IS*1272*, IS*1548*-like, IS*4*-like, IS*630*-SpnII, and IS*66* elements), and was larger than the others (Fig. 5). Indeed, 246 genes are exclusive to the LMG 17956 ST28 strain (114 of these 246 genes had been described only in *Streptococcus pneumoniae* and *Streptococcus pyogenes*) (Fig. 6).

The likelihood framework provided by BadiRate allows testing of whether these LMG 17956 ST28 gene acquisitions can be explained merely by the stochasticity underlying the mutational process or whether they have some selective meaning. Remarkably, our analyses identified some orthologous groups that were significantly expanded in the LMG 17956 ST28 lineage, including two related to transposable elements (an IS*110* element and a recombinase), as well as a few associated with antibiotic resistance {streptogramin A acetyltransferase, virginiamycin lyase, the 6′-aminoglycoside nucleotidyltransferase [ANT(6′)], and a complete system for bacteriocin production} (Table 4).

DISCUSSION

For humans, aside from its involvement in endocarditis, the major clinical interest of *S. gallolyticus* subsp. *gallolyticus* is the association with colorectal cancer. A new "bacterial driver-passenger model" has been proposed to explain the role of the gut microbiota in the colorectal cancer process (3). This model differentiates between the "bacterial drivers" in the microbiota (essentially the

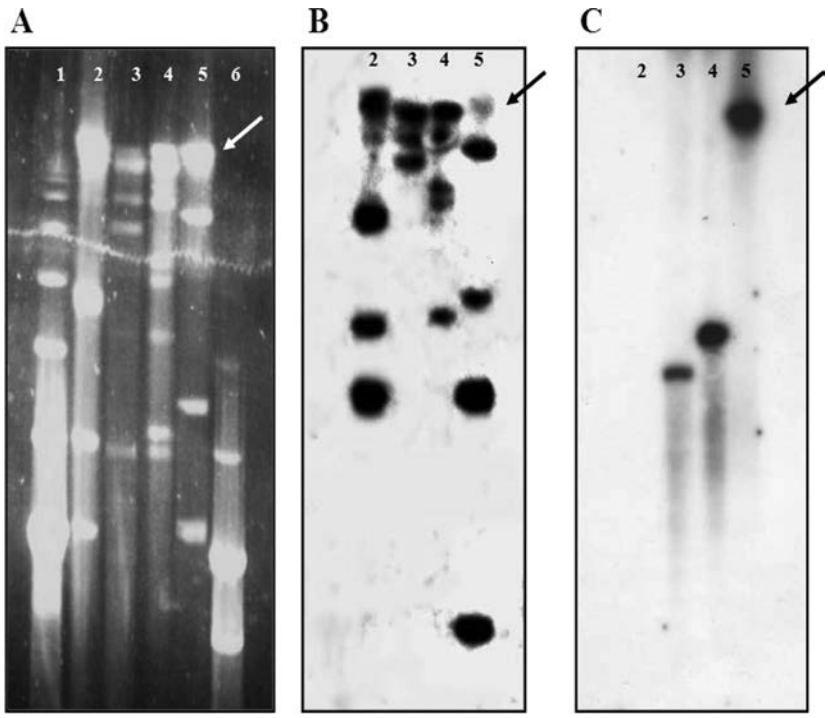


FIG 2 Hybridization experiments performed to determine if the Tn*1546*-like element is located in plasmids or on the bacterial chromosome. (A) CeuI digestion of the total DNAs of a negative-control vancomycin-susceptible *S. gallolyticus* subsp. *gallolyticus* strain (lane 2), positive-control *vanA*-containing *Enterococcus faecalis* (lane 3) and *Enterococcus faecium* (lane 4) strains, and the *vanA*-containing *S. gallolyticus* subsp. *gallolyticus* LMG 17956 ST28 strain (lane 5). Lanes 1 and 6 correspond to lambda and low-range markers. (B) Hybridization with the 16S rRNA genes of the CeuI fragments. The positive bands correspond to chromosomal fragments. (C) Hybridization with the *vanA* gene. Arrows point to the Tn*1546*-like element, which was integrated into the bacterial chromosome.

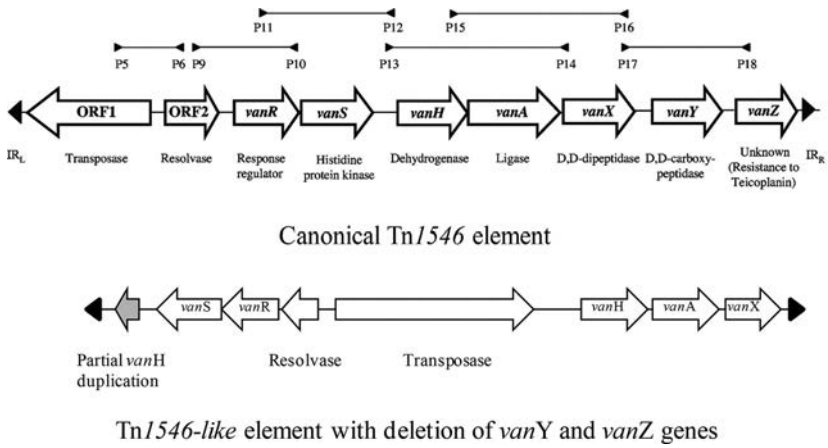


FIG 3 Genetic map of the Tn*1546*-like element detected in the LMG 17956 ST28 strain in comparison with the map of canonical Tn*1546*. The primers used in the primer-walking strategy are indicated above the map of the canonical Tn*1546* element (25). At the right of the *vanA* and *vanX* genes, the *vanY-vanZ* region present on the consensus Tn*1546* is absent. Note the partial duplication of the *vanH* gene.

genus *Bacteroides* and the family *Enterobacteriaceae*), which are directly related to the process of carcinogenesis, and the tumor-foraging opportunistic pathogens called "bacterial passengers."

The *S. bovis* complex belongs to the bacterial passenger category, mainly as a consequence of its efficient metabolic process, which permits adaptation to adverse environments. The prevalence of colonization of healthy human populations and animals has scarcely been investigated (8), and the presence of these organisms might be underestimated by a low fecal load.

The recent changes in the taxonomy of the *S. bovis* complex preclude a reliable retrospective analysis of most of the published epidemiological studies, except those that have investigated genetic diversity by the PFGE technique (38, 39). The multilocus sequence typing (MLST) tool has recently been adapted for the *S.*

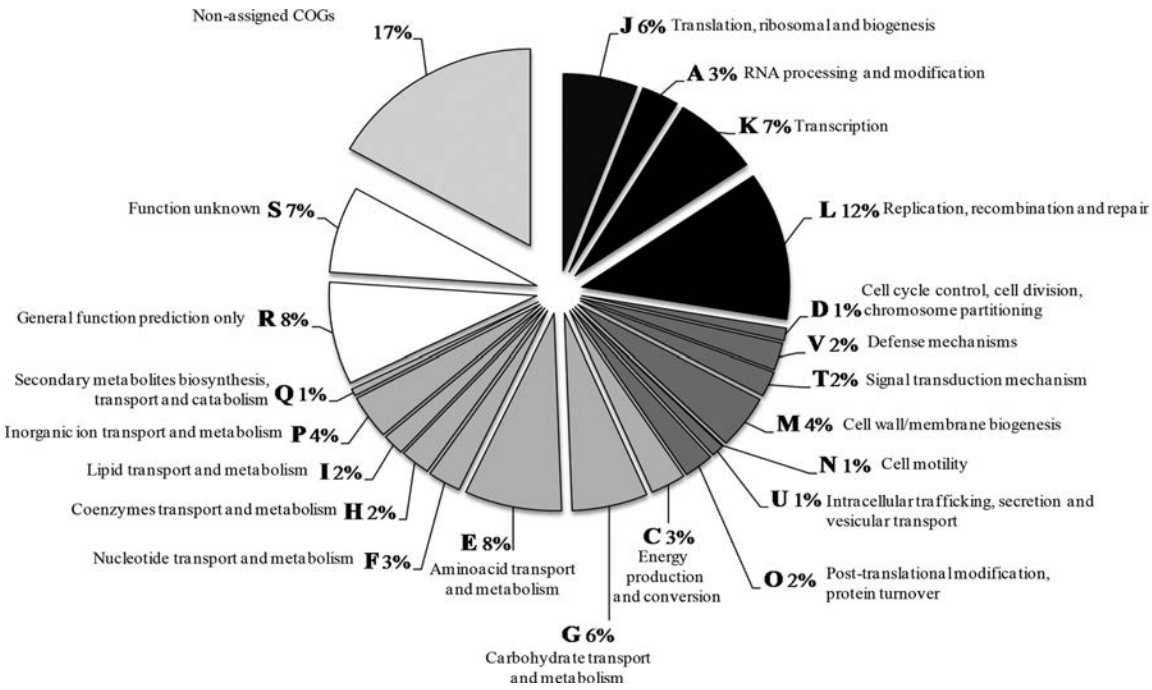


FIG 4 Distribution of COGs in the whole-genome sequence of the LMG 17956 ST28 isolate.

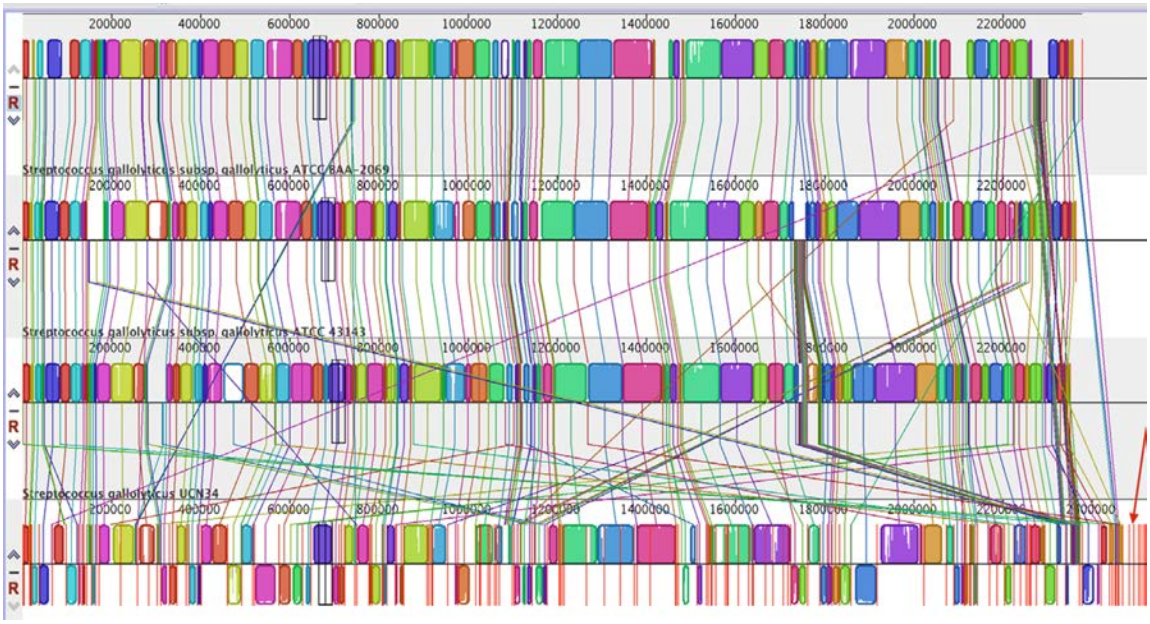


FIG 5 MAUVE comparison of the genome sequences of the *S. gallolyticus* subsp. *gallolyticus* ATCC BAA-2069, ATCC 43143, and UCN34 strains with that of the vancomycin-resistant LMG 17956 ST28 isolate. Note the extensive rearrangements in our isolate. The arrow points to the Tn*1546*-like element, and the red lines in the distal section correspond to transposases and IS.

gallolyticus subsp. *gallolyticus* population, enabling investigators to start deciphering the genetic population structure of this organism (10). In our work, the PFGE patterns of the animal strains (group I) were clearly separated from those of strains causing bacteremic episodes in humans (groups II, III, and IV); moreover, the high genetic diversity among our isolates was corroborated by MLST. Nevertheless, it should be noted that the vancomycin-resistant strain LMG 17956 ST28 recovered from a calf was grouped with the human-invasive isolates. This fact implies a risk for introduction of a vancomycin-resistant isolate into human compartments, particularly in view of the genetic plasticity of this strain, which could facilitate the acquisition of new resistant/virulent determinants.

The only published study focusing on genetic differences between animal and human isolates failed to distinguish between strains of the two collections by use of the randomly amplified polymorphic DNA (RAPD) and amplified rRNA gene restriction analysis (ARDRA) techniques (40). Currently, PFGE remains the gold standard technique for exploring genetic diversity in epidemiological studies, and differences between the work of Sasaki et al. (40) and our work are probably the consequence of the different methodologies used. Hence, the whole-genome sequencing strategy is the best option for objective comparison of results between laboratories. In spite of the particular characteristics of each origin, the existence of host-adapted genetic lineages, as in the *E. faecium* population, cannot be ruled out (41).

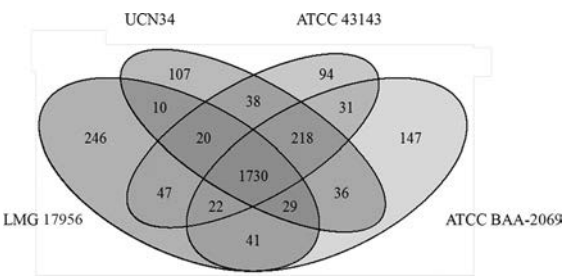


FIG 6 Venn diagram showing the exclusive and shared ortholog groups defined by OrthoMCL for the 4 *S. gallolyticus* genomes.

In general, *S. gallolyticus* subsp. *gallolyticus* is susceptible to antimicrobial compounds frequently used for humans. However, previous reports have described resistance to macrolides (45 to 59%) and tetracyclines (56 to 78%) and high-level resistance to aminoglycosides (35 to 43%) (1, 14–16, 42). Therefore, the therapy of *S. gallolyticus* subsp. *gallolyticus* infections with the standard penicillin antibiotic regime is uncomplicated in the clinical routine, since resistant strains have not yet been described.

In our work, we detected a glycopeptide-resistant isolate recovered from the feces of a calf. Similar reports in the literature are limited and involve strains of both animal and human origins, clinical and colonization sources, and the enterococcal genetic mechanisms *vanA* and *vanB* (19–21). The gut microbiota is a complex ecosystem, with coexistence, and possibly colocalization, of bacterial organisms, which are frequently genetically related and able to exchange genetic material (genetic exchange communities) (43). These communities might include both *Enterococcus* species and the *S. bovis* complex. Our hypothesis is that the

TABLE 4 Gene families significantly expanded in the *S. gallolyticus* subsp. *gallolyticus* LMG 17956 ST28 isolate

Molecular function	No. of genes in the following *S. gallolyticus* subsp. *gallolyticus* strain:				Statistical support[a]
	ATCC 43143	ATCC BAA-2069	UCN34	LMG 17956 ST28	
Transposase	0	1	0	7	28.789
Site-specific recombinase	0	0	0	4	28.621
Streptogramin A acetyltransferase	0	0	0	2	1.968
Virginiamycin B lyase (streptogramin B lyase)	0	0	0	2	1.968
Sequence-specific DNA binding	0	0	0	2	1.968
Dysgalacticin DysA2	0	0	0	2	1.968
UPF0236 protein in *vanSb* 3′ region	0	0	0	2	1.968
Uncharacterized protein	0	0	0	2	1.968
Uncharacterized protein	0	0	0	2	1.968

[a] The statistical support of these expansions was measured as the Akaike information criterion ratio between the first and the second best-fit models, as described in reference 32.

Tn*1546*-like element was originally located on an *Enterococcus* strain, was transmitted to *S. gallolyticus* subsp. *gallolyticus* by conjugation, and finally was integrated into the chromosome, provoking a deletion in the distal section.

The VanZ protein has been associated with teicoplanin resistance (44). Our isolate exhibited high-level resistance to both glycopeptides, even though the *vanY-vanZ* region was deleted. This result indicates that other pathways might contribute to the final inhibitory concentration of teicoplanin.

Although the primer-walking method failed in the characterization of the Tn*1546*-like element, whole-genome bacterial sequencing is an increasingly widely used strategy that allowed us to decode the structure of this transposon. In enterococci, differences between Tn*1546*-related elements from animals and humans (25, 45), as well as heterogeneity in Tn*1546*-like elements with IS*1216* insertions (46, 47), have been described previously. These features has not been reported previously for *S. gallolyticus* subsp. *gallolyticus*. In any case, the presence of a Tn*1546*-like element in an *S. gallolyticus* subsp. *gallolyticus* strain genetically close to those causing bacteremia and endocarditis in humans is certainly a matter of concern due to the zoonotic transmission possibility.

The comparative genomic analyses demonstrated that the LMG 17956 ST28 strain is quite different from the three described previously (48–50), due mainly to the presence of numerous insertion elements (IS) and other transposable elements. This unusual large number of transposable elements could provide considerable genome plasticity and might explain the large size of this strain's genome. Indeed, this strain contains 246 exclusive genes, some of which show statistical support for a lineage-specific gene family expansion. Our results are in agreement with those of Richards et al. (51), whose findings show a higher number of gene gains for bovine than for human isolates of *Streptococcus agalactiae*. These accelerations of the bovine gene gain rates probably reflect the adaptive pressure in the calf gut environment, where the microbiome diversity might facilitate horizontal gene transfer events, especially the transfer of antibiotic resistances and transposable elements.

ACKNOWLEDGMENTS

We are grateful to Dennis Hinse and Jens Dreier of the Institut für Laboratoriums- und Transfusionsmedizin, Herz- und Diabeteszentrum Nordrhein-Westfalen, Universitätsklinik der Ruhr-Universität Bochum, in Bad Oeynhausen, Germany, for sharing their bacterial collection. Technical support from Ana Moreno is also appreciated.

B.R.-H. has a "Rio Hortega" (CM11/181) contract from the Instituto de Salud Carlos III-FIS. This study was supported by Plan Nacional de I+D+i 2008-2011 and Instituto de Salud Carlos III, Subdirección General de Redes y Centros de Investigación Cooperativa, Ministerio de Economía y Competitividad, Spanish Network for Research in Infectious Diseases (REIPI RD12/0015). The study was cofinanced by the European Commission Project EvoTAR-282004 and the Fondo de Investigación Sanitaria (grant PI13/0205).

We have no conflict of interest to declare.

REFERENCES

1. **Romero B, Morosini MI, Loza E, Rodríguez-Baños M, Navas E, Cantón R, Campo RD.** 2011. Reidentification of *Streptococcus bovis* isolates causing bacteremia according to the new taxonomy criteria: still an issue? J Clin Microbiol **49:**3228–3233. http://dx.doi.org/10.1128/JCM.00524-11.
2. **Lazarovitch T, Shango M, Levine M, Brusovansky R, Akins R, Hayakawa K, Lephart PR, Sobel JD, Kaye KS, Marchaim D.** 2013. The relationship between the new taxonomy of *Streptococcus bovis* and its clonality to colon cancer, endocarditis, and biliary disease. Infection **41:**329–337. http://dx.doi.org/10.1007/s15010-012-0314-x.
3. **Boleij A, Tjalsma H.** 2013. The itinerary of *Streptococcus gallolyticus* infection in patients with colonic malignant disease. Lancet Infect Dis **13:**719–724. http://dx.doi.org/10.1016/S1473-3099(13)70107-5.
4. **Noble CJ.** 1978. Carriage of group D streptococci in the human bowel. J Clin Pathol **31:**1182–1186. http://dx.doi.org/10.1136/jcp.31.12.1182.
5. **Hedegaard L, Christensen H, Chadfield MS, Christensen JP, Bisgaard M.** 2009. Association of *Streptococcus pluranimalium* with valvular endocarditis and septicaemia in adult broiler parents. Avian Pathol **38:**155–160. http://dx.doi.org/10.1080/03079450902737763.
6. **Sekizaki T, Nishiya H, Nakajima S, Nishizono M, Kawano M, Okura M, Takamatsu D, Nishino H, Ishiji T, Osawa R.** 2008. Endocarditis in chickens caused by subclinical infection of *Streptococcus gallolyticus* subsp. *gallolyticus*. Avian Dis **52:**183–186. http://dx.doi.org/10.1637/8048-070307-Case.
7. **Chadfield MS, Christensen JP, Decostere A, Christensen H, Bisgaard M.** 2007. Geno- and phenotypic diversity of avian isolates of *Streptococcus gallolyticus* subsp. *gallolyticus* (*Streptococcus bovis*) and associated diagnostic problems. J Clin Microbiol **45:**822–827. http://dx.doi.org/10.1128/JCM.00922-06.
8. **Devriese LA, Vandamme P, Pot B, Vanrobaeys M, Kersters K, Haesebrouck F.** 1998. Differentiation between *Streptococcus gallolyticus* strains of human clinical and veterinary origins and *Streptococcus bovis* strains from the intestinal tracts of ruminants. J Clin Microbiol **36:**3520–3523.
9. **Garvie EI, Bramley AJ.** 1979. *Streptococcus bovis*—an approach to its classification and its importance as a cause of bovine mastitis. J Appl Bacteriol **46:**557–566. http://dx.doi.org/10.1111/j.1365-2672.1979.tb00855.x.
10. **Dumke J, Hinse D, Vollmer T, Knabbe C, Dreier J.** 2014. Development and application of a multilocus sequence typing scheme for *Streptococcus gallolyticus* subsp. *gallolyticus*. J Clin Microbiol **52:**2472–2478. http://dx.doi.org/10.1128/JCM.03329-13.

11. **Danne C, Entenza JM, Mallet A, Briandet R, Débarbouillé M, Nato F, Glaser P, Jouvion G, Moreillon P, Trieu-Cuot P, Dramsi S.** 2011. Molecular characterization of a *Streptococcus gallolyticus* genomic island encoding a pilus involved in endocarditis. J Infect Dis **204:**1960–1970. http://dx.doi.org/10.1093/infdis/jir666.
12. **Sillanpää J, Nallapareddy SR, Qin X, Singh KV, Muzny DM, Kovar CL, Nazareth LV, Gibbs RA, Ferraro MJ, Steckelberg JM, Weinstock GM, Murray BE.** 2009. A collagen-binding adhesin, Acb, and ten other putative MSCRAMM and pilus family proteins of *Streptococcus gallolyticus* subsp. *gallolyticus* (*Streptococcus bovis* group, biotype I). J Bacteriol **191:**6643–6653. http://dx.doi.org/10.1128/JB.00909-09.
13. **Vollmer T, Hinse D, Kleesiek K, Dreier J.** 2010. Interactions between endocarditis-derived *Streptococcus gallolyticus* subsp. *gallolyticus* isolates and human endothelial cells. BMC Microbiol **10:**78. http://dx.doi.org/10.1186/1471-2180-10-78.
14. **Kimpe A, Decostere A, Martel A, Devriese LA, Haesebrouck F.** 2003. Phenotypic and genetic characterization of resistance against macrolides and lincosamides in *Streptococcus gallolyticus* strains isolated from pigeons and humans. Microb Drug Resist **9**(Suppl 1):S35–S38. http://dx.doi.org/10.1089/107662903322541874.
15. **Leclercq R, Huet C, Picherot M, Trieu-Cuot P, Poyart C.** 2005. Genetic basis of antibiotic resistance in clinical isolates of *Streptococcus gallolyticus* (*Streptococcus bovis*). Antimicrob Agents Chemother **49:**1646–1648. http://dx.doi.org/10.1128/AAC.49.4.1646-1648.2005.
16. **Nomoto R, Tien LHT, Sekizaki T, Osawa R.** 2013. Antimicrobial susceptibility of *Streptococcus gallolyticus* isolated from humans and animals. Jpn J Infect Dis **66:**334–336. http://dx.doi.org/10.7883/yoken.66.334.
17. **Dahl KH, Sundsfjord A.** 2003. Transferable *vanB2* Tn*5382*-containing elements in fecal streptococcal strains from veal calves. Antimicrob Agents Chemother **47:**2579–2583. http://dx.doi.org/10.1128/AAC.47.8.2579-2583.2003.
18. **Mevius D, Devriese L, Butaye P, Vandamme P, Verschure M, Veldman K.** 1998. Isolation of glycopeptide resistant *Streptococcus gallolyticus* strains with *vanA*, *vanB*, and both *vanA* and *vanB* genotypes from faecal samples of veal calves in The Netherlands. J Antimicrob Chemother **42:**275–276. http://dx.doi.org/10.1093/jac/42.2.275.
19. **Poyart C, Pierre C, Quesne G, Pron B, Berche P, Trieu-Cuot P.** 1997. Emergence of vancomycin resistance in the genus *Streptococcus*: characterization of a *vanB* transferable determinant in *Streptococcus bovis*. Antimicrob Agents Chemother **41:**24–29.
20. **Corredoira J, Alonso MP, Pita J, Alonso-Mesonero D.** 2008. Association between rural residency, group D streptococcal endocarditis and colon cancer? Clin Microbiol Infect **14:**190. http://dx.doi.org/10.1111/j.1469-0691.2007.01913.x.
21. **Giannitsioti E, Chirouze C, Bouvet A, Béguinot I, Delahaye F, Mainardi JL, Celard M, Mihaila-Amrouche L, Moing VL, Hoen B; AEPEI Study Group.** 2007. Characteristics and regional variations of group D streptococcal endocarditis in France. Clin Microbiol Infect **13:**770–776. http://dx.doi.org/10.1111/j.1469-0691.2007.01753.x.
22. **Hinse D, Vollmer T, Erhard M, Welker M, Moore ER, Kleesiek K, Dreier J.** 2011. Differentiation of species of the *Streptococcus bovis/equinus*-complex by MALDI-TOF mass spectrometry in comparison to *sodA* sequence analyses. Syst Appl Microbiol **34:**52–57. http://dx.doi.org/10.1016/j.syapm.2010.11.010.
23. **Clinical and Laboratory Standards Institute.** 2007. Performance standards for antimicrobial susceptibility testing; 17th informational supplement. Document M100-S17. Clinical and Laboratory Standards Institute, Wayne, PA.
24. **Luey CK, Chu YW, Cheung TK, Law CC, Chu MY, Cheung DT, Kam KM.** 2007. Rapid pulsed-field gel electrophoresis protocol for subtyping of *Streptococcus suis* serotype 2. J Microbiol Methods **68:**648–650. http://dx.doi.org/10.1016/j.mimet.2006.10.010.
25. **Woodford N, Adebiyi AM, Palepou MF, Cookson BD.** 1998. Diversity of VanA glycopeptide resistance elements in enterococci from humans and nonhuman sources. Antimicrob Agents Chemother **42:**502–508.
26. **Chevreux B, Wetter T, Suhai S.** 1999. Genome sequence assembly using trace signals and additional sequence information, p 45–56. *In* Wingender E (ed), Computer science and biology: proceedings of the German Conference on Bioinformatics (GCB '99). Gesellschaft fur Biotechnologische Forschung, Department of Bioinformatics, Braunschweig, Germany.
27. **Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA.** 2008. Database indexing for production MegaBLAST searches. Bioinformatics **24:**1757–1764. http://dx.doi.org/10.1093/bioinformatics/btn322.
28. **Zhang Z, Schwartz S, Wagner L, Miller W.** 2000. A greedy algorithm for aligning DNA sequences. J Comput Biol **7:**203–214. http://dx.doi.org/10.1089/10665270050081478.
29. **Darling AE, Mau B, Perna NT.** 2010. progressiveMauve: multiple genome alignment with gene gain, loss, and rearrangement. PLoS One **5:**e11147. http://dx.doi.org/10.1371/journal.pone.0011147.
30. **Li L, Stoeckert CJ, Jr, Roos DS.** 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res **13:**2178–2189. http://dx.doi.org/10.1101/gr.1224503.
31. **Fischer S, Brunk BP, Chen F, Gao X, Harb OS, Iodice JB, Shanmugam D, Roos DS, Stoeckert CJ, Jr.** 2011. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. Curr Protoc Bioinformatics Chapter 6:Unit 6.12.1–19. http://dx.doi.org/10.1002/0471250953.bi0612s35.
32. **Chen H, Boutros PC.** 2011. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinformatics **12:**35. http://dx.doi.org/10.1186/1471-2105-12-35.
33. **Librado P, Vieira FG, Rozas J.** 2012. BadiRate: estimating family turnover rates by likelihood-based methods. Bioinformatics **28:**279–281. http://dx.doi.org/10.1093/bioinformatics/btr623.
34. **Katoh K, Standley DM.** 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol **30:**772–780. http://dx.doi.org/10.1093/molbev/mst010.
35. **Stamatakis A.** 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22:**2688–2690. http://dx.doi.org/10.1093/bioinformatics/btl446.
36. **Stamatakis A, Ott M.** 2008. Efficient computation of the phylogenetic likelihood function on multi-gene alignments and multi-core architectures. Philos Trans R Soc Lond B Biol Sci **363:**3977–3984. http://dx.doi.org/10.1098/rstb.2008.0163.
37. **Arthur M, Molinas C, Depardieu F, Courvalin P.** 1993. Characterization of Tn*1546*, a Tn*3*-related transposon conferring glycopeptide resistance by synthesis of depsipeptide peptidoglycan precursors in *Enterococcus faecium* BM4147. J Bacteriol **175:**117–127.
38. **Tripodi MF, Fortunato R, Utili R, Triassi M, Zarrilli R.** 2005. Molecular epidemiology of *Streptococcus bovis* causing endocarditis and bacteraemia in Italian patients. Clin Microbiol Infect **11:**814–819. http://dx.doi.org/10.1111/j.1469-0691.2005.01248.x.
39. **Wang SM, Deighton MA, Capstick JA, Gerraty N.** 1999. Epidemiological typing of bovine streptococci by pulsed-field gel electrophoresis. Epidemiol Infect **123:**317–324. http://dx.doi.org/10.1017/S0950268899002745.
40. **Sasaki E, Osawa R, Nishitani Y, Whiley RA.** 2004. ARDRA and RAPD analyses of human and animal isolates of *Streptococcus gallolyticus*. J Vet Med Sci **66:**1467–1470. http://dx.doi.org/10.1292/jvms.66.1467.
41. **Willems RJ, Top J, van den Braak N, van Belkum A, Endtz H, Mevius D, Stobberingh E, van den Bogaard A, van Embden JD.** 2000. Host specificity of vancomycin-resistant *Enterococcus faecium*. J Infect Dis **182:**816–823. http://dx.doi.org/10.1086/315752.
42. **Rodríguez-Avial I, Rodríguez-Avial C, Culebras E, Picazo JJ.** 2007. Fluoroquinolone resistance among invasive viridans group streptococci and *Streptococcus bovis* isolated in Spain. Int J Antimicrob Agents **29:**478–480. http://dx.doi.org/10.1016/j.ijantimicag.2006.12.006.
43. **Baquero F, Tedim AP, Coque TM.** 2013. Antibiotic resistance shaping multi-level population biology of bacteria. Front Microbiol **4:**15. http://dx.doi.org/10.3389/fmicb.2013.00015.
44. **Arthur M, Depardieu F, Molinas C, Reynolds P, Courvalin P.** 1995. The *vanZ* gene of Tn*1546* from *Enterococcus faecium* BM4147 confers resistance to teicoplanin. Gene **154:**87–92. http://dx.doi.org/10.1016/0378-1119(94)00851-I.
45. **Willems RJ, Top J, van den Braak N, van Belkum A, Mevius DJ, Hendriks G, van Santen-Verheuvel M, van Embden JD.** 1999. Molecular diversity and evolutionary relationships of Tn*1546*-like elements in enterococci from humans and animals. Antimicrob Agents Chemother **43:**483–491. http://dx.doi.org/10.1093/jac/43.4.483.
46. **Jensen LB.** 1998. Internal size variations in Tn*1546*-like elements due to the presence of IS*1216*V. FEMS Microbiol Lett **169:**349–354. http://dx.doi.org/10.1111/j.1574-6968.1998.tb13339.x.
47. **Talebi M, Pourshafie MR, Katouli M, Möllby R.** 2008. Molecular structure and transferability of Tn*1546*-like elements in *Enterococcus faecium* isolates from clinical, sewage, and surface water samples in Iran. Appl Environ Microbiol **74:**1350–1356. http://dx.doi.org/10.1128/AEM.02254-07.
48. **Hinse D, Vollmer T, Rückert C, Blom J, Kalinowski J, Knabbe C, Dreier J.** 2011. Complete genome and comparative analysis of *Streptococcus gal-*

lolyticus subsp. *gallolyticus*, an emerging pathogen of infective endocarditis. BMC Genomics **12:**400. http://dx.doi.org/10.1186/1471-2164-12-400.

49. **Lin IH, Liu TT, Teng YT, Wu HL, Liu YM, Wu KM, Chang CH, Hsu MT.** 2011. Sequencing and comparative genome analysis of two pathogenic *Streptococcus gallolyticus* subspecies: genome plasticity, adaptation and virulence. PLoS One **6:**e20519. http://dx.doi.org/10.1371/journal.pone.0020519.
50. **Rusniok C, Couvé E, Da Cunha V, El Gana R, Zidane N, Bouchier C, Poyart C, Leclercq R, Trieu-Cuot P, Glaser P.** 2010. Genome sequence of *Streptococcus gallolyticus*: insights into its adaptation to the bovine rumen and its ability to cause endocarditis. J Bacteriol **192:**2266–2276. http://dx.doi.org/10.1128/JB.01659-09.
51. **Richards VP, Palmer SR, Pavinski Bitar PD, Qin X, Weinstock GM, Highlander SK, Town CD, Burne RA, Stanhope MJ.** 2014. Phylogenomics and the dynamic genome evolution of the genus *Streptococcus*. Genome Biol Evol **6:**741–753. http://dx.doi.org/10.1093/gbe/evu048.

A.2. Gene duplications in the *E.coli* genome: common themes among pathotypes.

Las duplicaciones génicas suponen una proporción importante de diversidad funcional y complejidad genómica tanto en eucariotas como en procariotas. A pesar de que encontramos descritos en la literatura algunos genes duplicados de *Escherichia coli*, es necesario un análisis extensivo de las duplicaciones génicas de este organismo. Los genomas de cepas como la *E.coli* enteroagregativa 042 y otras cepas patógenas contienen duplicaciones génicas que codifican para un regulador global denominado *hha*. Para determinar si la presencia de copias adicionales del gen *hha* se correlacionan con la presencia de otros genes, desarrollamos un análisis genómico comparado entre diferentes cepas de *E.coli* con y sin duplicaciones en *hha*. Los resultados muestran que cepas que presentan copias adicionales del gen *hha* también codifican el cluster *yeeR irmA* (*aec69*), que, a su vez, está también duplicado en esta cepa 042 y algunas otras cepas. La identificación de estas duplicaciones dio pie a la obtención de un mapa global de las duplicaciones génicas en la cepa 042 y en otros genomas de *E.coli*.

Mediante una búsqueda de similitud de secuencias de proteínas, BLASTP, se ha identificado las duplicaciones génicas descritas en los genomas de la cepa enteroagregativa 042, la cepa uropatogénica CFT073 y la cepa enterohemorrágica O145:H28. Esta metodología también fue empleada en la determinación de la distribución de duplicados idénticos entre genomas de un conjunto de 28 cepas representativas de *E.coli*. A pesar de la diversidad genómica de las cepas, hemos identificado múltiples duplicados en los genomas de casi todas las cepas patogénicas estudiadas. Muchos de los genes duplicados no tienen una función conocida. Análisis transcriptómicos también mostraron que muchos de los genes duplicados están reguladas por las proteinas H-NS/Hha.

Existen múltiples genes duplicados que están ampliamente distribuidos en cepas patogénicas de *E.coli*. Además, algunos de estos genes están duplicados en tan solo algunos patotipos y otros son específicos de una cepa determinada. Este análisis de duplicación génica muestra una relación entre patotipos de *E.coli* y sugiere que estos nuevos genes duplicados identificados en una alta cantidad de cepas patógenas puede tener un papel relevante en la virulencia. Nuestro estudio también muestra la relación entre las duplicaciones génicas de genes reguladores y la de genes codificados por sus dianas de regulación.

A

Bernabeu et al. BMC Genomics (2019) 20:313
https://doi.org/10.1186/s12864-019-5683-4

BMC Genomics

RESEARCH ARTICLE Open Access

Gene duplications in the *E. coli* genome: common themes among pathotypes

Manuel Bernabeu[1], José Francisco Sánchez-Herrero[1,2], Pol Huedo[3], Alejandro Prieto[1], Mário Hüttener[1], Julio Rozas[1,2] and Antonio Juárez[1,4*]

Abstract

Background: Gene duplication underlies a significant proportion of gene functional diversity and genome complexity in both eukaryotes and prokaryotes. Although several reports in the literature described the duplication of specific genes in *E. coli*, a detailed analysis of the extent of gene duplications in this microorganism is needed.

Results: The genomes of the *E. coli* enteroaggregative strain 042 and other pathogenic strains contain duplications of the gene that codes for the global regulator Hha. To determine whether the presence of additional copies of the *hha* gene correlates with the presence of other genes, we performed a comparative genomic analysis between *E. coli* strains with and without *hha* duplications. The results showed that strains harboring additional copies of the *hha* gene also encode the *yeeR irmA* (*aec69*) gene cluster, which, in turn, is also duplicated in strain 042 and several other strains. The identification of these duplications prompted us to obtain a global map of gene duplications, first in strain 042 and later in other *E. coli* genomes.
Duplications in the genomes of the enteroaggregative strain 042, the uropathogenic strain CFT073 and the enterohemorrhagic strain O145:H28 have been identified by a BLASTp protein similarity search. This algorithm was also used to evaluate the distribution of the identified duplicates among the genomes of a set of 28 representative *E. coli* strains. Despite the high genomic diversity of *E. coli* strains, we identified several duplicates in the genomes of almost all studied pathogenic strains. Most duplicated genes have no known function. Transcriptomic analysis also showed that most of these duplications are regulated by the H-NS/Hha proteins.

Conclusions: Several duplicated genes are widely distributed among pathogenic *E. coli* strains. In addition, some duplicated genes are present only in specific pathotypes, and others are strain specific. This gene duplication analysis shows novel relationships between *E. coli* pathotypes and suggests that newly identified genes that are duplicated in a high percentage of pathogenic *E. coli* isolates may play a role in virulence. Our study also shows a relationship between the duplication of genes encoding regulators and genes encoding their targets.

Keywords: Pathotypes, Gene duplication, *Escherichia coli* 042, H-NS, Hha

Background

Pathogenic *Escherichia coli* strains can cause either intestinal infections (which are diarrheagenic) or extraintestinal infections. Based on the type of virulence factors displayed and the strategy used to cause infection, *E. coli* strains are grouped into pathotypes. Some pathotypes are associated with diarrhea: enteropathogenic (EPEC), enterotoxigenic (ETEC), enterohemorrhagic (EHEC), enteroaggregative (EAEC) and enteroinvasive (EIEC) strains are the best characterized. Other pathotypes are common causes of urinary tract infections (uropathogenic *E. coli*, UPEC), newborn meningitis (neonatal meningitis *E. coli*, NMEC) or sepsis (SEPEC).

As mentioned above, enteroaggregative *Escherichia coli* (EAEC) strains are one of the groups of diarrheal *E. coli* pathogens [1]. EAEC strains can be distinguished from EPEC strains because of their different patterns of adherence to HEp-2 cells. Whereas EPEC strains display a "microcolony" pattern of adherence, EAEC strains display a characteristic aggregative or "stacked-brick" pattern [2].

* Correspondence: ajuarez@ub.edu
[1]Department of Genetics, Microbiology and Statistics, University of Barcelona, Barcelona, Spain
[4]Institute for Bioengineering of Catalonia, The Barcelona Institute of Science and Technology, Barcelona, Spain
Full list of author information is available at the end of the article

EAEC adherence to intestinal cells is mediated by a specific type of fimbrial adhesin termed aggregative adherence fimbriae (AAFs). Epidemiological studies have shown that EAEC strains are genetically heterogeneous. A large number of virulence factors have been identified in EAEC clinical isolates [3]. Most of these virulence factors are expressed by only a percentage of all EAEC strains characterized. The strain O104:H4 is an example of EAEC genetic heterogeneity. A few years ago in Germany, this strain caused a large outbreak of bloody diarrhea [4]. Isolates from the O104:H4 outbreak harbor a plasmid (pAA) that encodes, among other virulence factors, the fimbriae that mediate the EAEC type of adherence [5]. Unlike typical EAEC strains, strain O104:H4 contains a prophage encoding the Shiga toxin [6], which is a well-characterized virulence determinant usually expressed by a different *E. coli* pathotype, EHEC [7].

Strain 042 is the best-characterized EAEC strain. It caused diarrhea in a volunteer trial [8]. The genome sequence of this strain is available [9], and its virulence factors are characterized. Strain 042 harbors the IncFIC virulence plasmid pAA2 [9, 10], which encodes the fimbrial adhesion determinants (the AAF/II variant of AAF), the virulence master regulator AggR and other virulence determinants [9, 11–13].

When analyzing the 042 genomic sequence, we noticed that unlike other *E. coli* strains, the chromosome of this strain encodes four paralogues of the *hha* gene: *hha, ydgT* and the novel *hha2* and *hha3* genes [14]. The *hha* gene product, the Hha protein, is representative of a family that includes a group of sequence-related low molecular mass proteins (approximately 8 kDa) involved in gene regulation in enterobacteria. These proteins interact with the nucleoid-associated protein H-NS to modulate gene expression (as reviewed in [15]). The genomes of several enterobacterial isolates, such as *Salmonella* and *E. coli* strains, encode a paralogue of the *hha* gene (the *ydgT* gene). Orthologues of *hha* are also present in several conjugative plasmids [16, 17]. The presence of the novel chromosomal *hha* paralogues *hha2* and *hha3* has been associated with pathogenic *E. coli* strains that belong to a wide range of pathotypes [14].

Gene duplication underlies a significant proportion of gene functional diversity and genome complexity [18–22]. Gene duplications occur in both eukaryotes and prokaryotes and significantly impact their gene repertoires [18–23]. In this work, we first aimed to gain insight into the biological role of the novel *hha2* and *hha3* genes of strain 042. To this end, we first performed a comparative genomic analysis between strains with and without *hha2/hha3*. This approach allowed us to correlate *hha2/hha3* with a gene cluster (the *flu yeeR* gene cluster), which is also duplicated in strain 042. Because strain 042 exhibits the duplication of genes encoding both regulators and the genes likely targeted by regulators, we decided to determine the extent of gene duplications in this strain and in the genomes of other pathogenic *E. coli* strains. Our analysis uncovers interesting patterns of gene duplications that are common to strains belonging to several *E. coli* pathotypes, both diarrheagenic and nondiarrheagenic.

Methods

To investigate the pan-, core, variable, and exclusive genomes of *E. coli hha*$^+$ (*hha2/3*$^+$) and *hha*$^-$ strains, two sets of five representative strains were considered. The *E. coli* strains in the *hha2/3*$^+$ set were 042, NA114, O104:H4 LB226692, ETEC H10407 and UMN026. The *E. coli* strains in the *hha*$^-$set were O111:H-11128, 53,638, IAI39, O127:H6 E2348/69 and O157:H7 Sakai (see Additional file 1: Table S1 for details).

Genomic analyses were performed using the MaGe Pan/Core genome tool (http://www.genoscope.cns.fr/agc/microscope/compgenomics/pancoreTool.php), and protein families were determined using MicroScope gene families (MICFAM) [24] with the following parameters: 80% amino acid identity and 80% alignment coverage.

For the identification of putative duplicates, we retrieved and downloaded the translated coding sequences of 28 *E. coli* strains from GenBank (Additional file 1: Table S1). For the BLAST search analysis, we used as filtering parameters a similarity cutoff > 85%, an alignment length between pairs > 85% and an *e*-value < 10^{-10}.

We analyzed the extent of gene duplication among strains by performing an all-vs-all BLASTp [25] protein similarity search (i.e., with the translated coding sequence regions of each strain, filtering the results according to the parameters specified above). For each duplicate, we retrieved genomic features (from the GenBank genomic feature files-gff), plotted the coordinates using R [26] and colored the duplicates according to their groups.

For the gene duplication analysis between strains and for the identification of the presence/absence of putative duplicated encoded proteins/coding regions, we also employed BLASTp. We searched the putative duplicates of interest against all translated coding sequences (all six frames). The results were filtered according to the above cutoff parameters.

In silico operon prediction was performed using the FGENESB program (Softberry, Inc., Mount Kisco, NY) (http://www.softberry.com/).

The bioinformatics scripts employed for the analysis were deposited and available at the github website: https://github.com/molevol-ub/BacterialDuplicates.

Statistical analysis. Proportions were compared between groups by using the two-tailed Fisher's exact test. A *P*-value of less than 0.05 was considered significant.

For the RNA-seq experiments, the detailed information and raw data were previously published in [27].

Results

E. coli strains encoding *hha2/hha3* usually encode the *flu yeeR aec69 aec70* cluster, which is also duplicated

To gain insight into the biological role of *hha* duplication in the EAEC strain 042, we hypothesized that the presence of multiple alleles of a global regulator could be associated with the presence of genes specifically targeted by the regulator. To support this hypothesis, we decided to compare the core genomes of two groups, each with five *E. coli* strains. One of them included representatives that encode *hha2/hha3* (hha^+), and the other included strains that do not encode them (hha^-). To identify those genes that are truly exclusive to the hha^+ set, we used a restrictive strategy of excluding the pangenome of the hha^- set from the core genome of the hha^+ set. By using this approach, only three gene families could be identified in the hha^+ set: the *hha*, *yeeR* and *aec69* genes (Additional file 1: Figure S1 and Additional file 2).

The yeeR and *aec69* genes belong to a gene cluster that includes *flu* (whose gene product is the well-characterized antigen43 protein), *aec70* and *aec71* (Fig. 1). A recent report shows that *aec69*, termed *irmA*, is transcribed in a single transcriptional unit with *flu* and *yeeR* [28]. In *E. coli* K12, the *yeeR* gene is truncated, and the *irmA* (*aec69*), *aec70* and *aec71* genes are missing (Fig. 1). This cluster belongs to the prophage CP4–44. Taking into account the high genomic variability of *E. coli,* the identification of *yeeR* and *irmA* (*aec69)* as linked to *hha2/hha3* when the two five-strain groups were compared does not exclude the possibility that other strains that do not encode *hha2/hha3* might encode *yeeR/irmA* (*aec69*) or that other strains harboring *hha2/hha3* do not harbor *yeeR/irmA* (*aec69*). To improve the analysis, we performed a BLASTp search on a total of 28 *E. coli* genomes, including both commensal and pathogenic strains belonging to several pathotypes (Additional file 1: Table S1). The results obtained showed that 72% of the strains expressing *hha2/hha3* also express the *yeeR* or *irmA* (*aec69*) genes (*P*-value < 0.05), while 61% also express *flu* or *aec70* (this latter comparison was close to the critical value ($P = 0.055$)) (Additional file 1: Table S2). In contrast, only 20–40% of $hha2/hha3^-$ strains express *yeeR, irmA (aec69), flu* and *aec70*. *aec71* does not appear to be associated with *hha2/hha3*. Its presence is widespread in both *hha2/hha3* and $hha2/hha3^-$ strains. We then analyzed the map positions of the *hha2/hha3* genes and the *flu yeeR irmA* (*aec69*) *aec70 aec71* gene cluster in the chromosomes of seven *E. coli* strains corresponding to different pathotypes, including both enteric and extraintestinal pathogens (Fig. 2). In several instances, *hha2/hha3* mapped close to the *yeeR irmA* (*aec69*) *aec70 aec71* gene cluster. This study also showed that in most of the virulent *E. coli* strains analyzed (including the EAEC strain 042), genes belonging to the *yeeR irmA* (*aec69*) *aec70 aec71* cluster are also duplicated (Fig. 2). The presence in the chromosome of strain 042 of four copies of *hha*--like genes (*hha, ydgT, hha2* and *hha3* [14]), three copies of *hns*-like genes (*hns, stpA* and *hns2*) [27], two copies of *yeeR* and *irmA* (*aec69*), three copies of *flu* and four copies of the *aec71* gene suggests that gene duplication may play a relevant role in this and perhaps other pathogenic *E. coli* strains. We therefore decided to investigate the extent of gene duplications first in the genome of strain 042 and thereafter in the genomes of other pathogenic *E. coli* strains.

Gene duplications in the EAEC strain 042 genome

We analyzed the extent of gene duplications in strain 042 by using the BLASTp algorithm (see the materials for details) and mapped along the 042 genome those genes that are present in two or more copies (Fig. 3a). A total of 80 genes were duplicated in strain 042. Some of these genes correspond to transposases (black and open circles). Most of the duplicated genes are clustered in

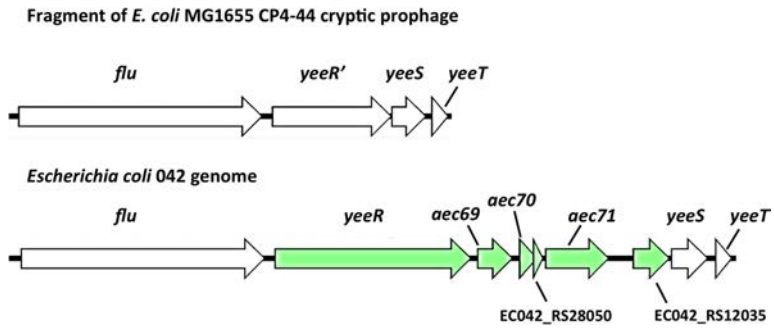


Fig. 1 *flu yeeR* gene cluster in *E. coli*. Schematic representation of the genetic cluster comprising the *flu, yeeR, yeeS and yeeT* genes in *E. coli* MG1655 prophage CP4–44 and *flu, yeeR, irmA* (*aec69*), *aec70, aec71, yeeS and yeeT* in *E. coli* 042

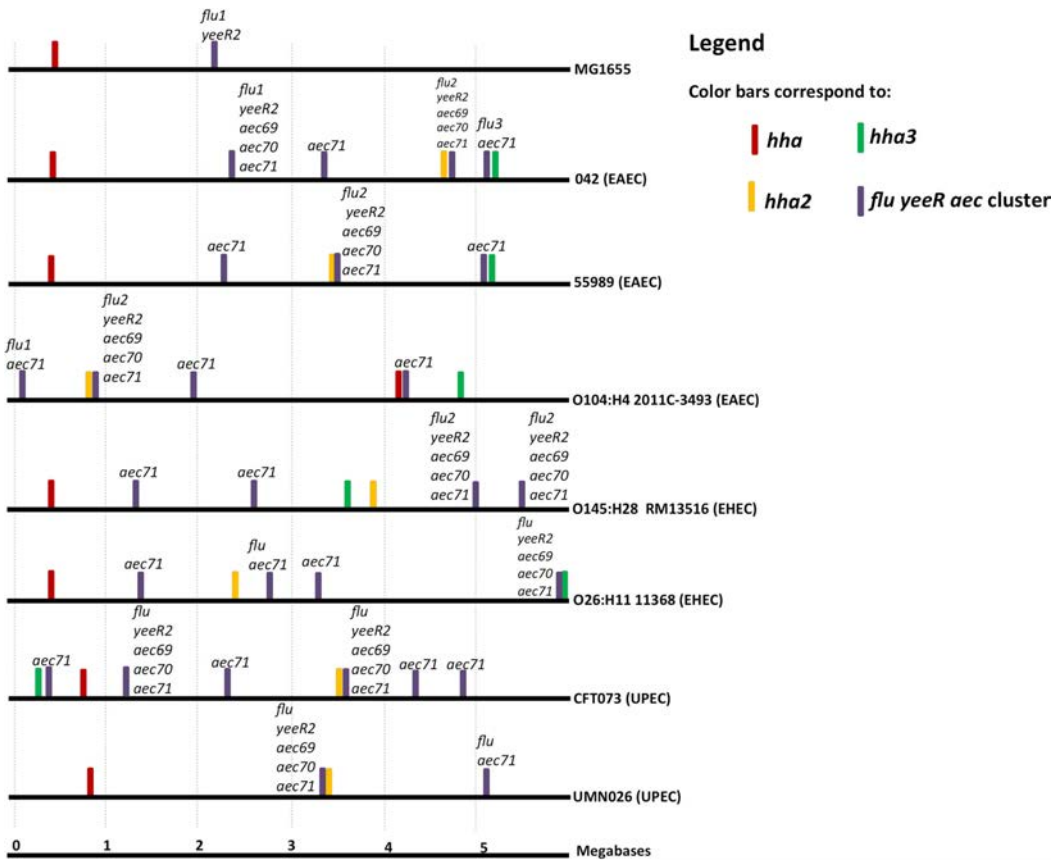


Fig. 2 Map position of the *hha* (red bar), *hha2* (orange bar), *hha3* (green bar) and genes from the *flu yeeR aec* gene cluster (violet bar) in the chromosomes of *E. coli* MG1655 and seven other *E. coli* strains belonging to different pathotypes (as indicated in each case). Note the duplication of genes belonging to the *flu yeeR aec* gene cluster in several strains. Scale bars 0 to 5 correspond to megabases in the genetic map

three main regions (labeled with vertical bars), which we arbitrarily termed regions 1 to 3. A significant number of genes that map to region 1, which is approximately 35.5 kb long, correspond to phage genes (Fig. 3b, Additional file 1: Table S3). Region 2 is approximately 17 kb long and contains the *flu yeeR irmA* (*aec69*) *aec70 aec71* cluster, a toxin-antitoxin gene and several other genes of unknown function (Fig. 3b, Additional file 1: Table S3). One of the copies of this region includes the *hha2* gene. The two copies of region 2 are inverted in the 042 chromosome, suggesting that genetic rearrangements leading to gene duplication can affect this region as a single recombinational unit. Region 3 is 10 kb long and includes mostly genes of unknown function (Fig. 3b, Additional file 1: Table S3).

We also analyzed gene duplications in strain 042 by using a BLASTn algorithm, yielding results similar to those obtained by using BLASTp (Additional file 1: Figure S2).

Duplicated genes in regions 1 and 2 are repressed by the H-NS/Hha system

Considering that some duplicated genes of region 2 in strain 042 (i.e., the *yeeR irmA* (*aec69*) gene cluster) have been identified as linked to *hha2/hha3*, it can be hypothesized that some duplicated genes are regulated by the H-NS/Hha system. To support this hypothesis, we analyzed the previously reported transcriptional profiles of strain 042 and its *hha* null (*hha hha2*) and *hns* mutant derivatives [27], which was performed in cultures growing in LB medium at 37 °C. We assessed whether the duplicated genes of the three regions of strain 042 show H-NS- or Hha-dependent regulation (Table 1). All genes from region 2 show fold change values higher than 2, both in the

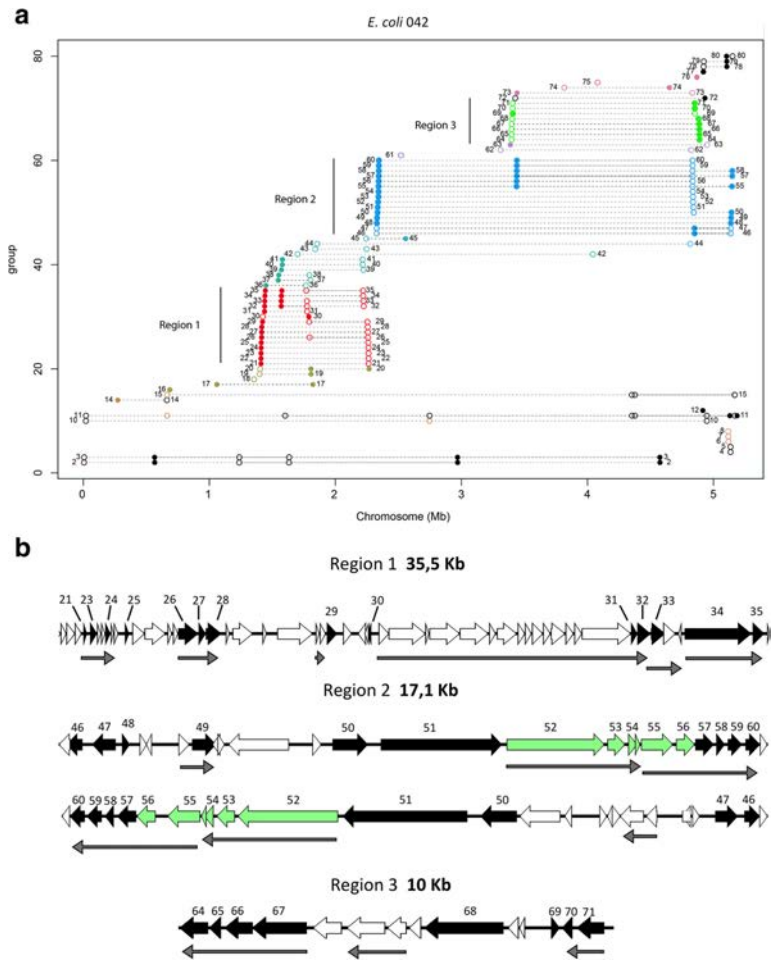


Fig. 3 a Genes duplicated in the *E. coli* strain 042. The X axis corresponds to the linear map of the chromosome. Each group of spots connected by a horizontal dashed line corresponds to a single gene duplicated in different positions on the chromosome. The different spots indicate the map positions of the different copies of the gene. Point shapes represent the strand on which a protein is encoded: filled circle for (+) strand and circle for (–) strand. Numbers correspond to the different duplicated genes, which have been numbered by their order starting from the origin of the chromosomal map. Genes numbered 2 to 5, 10 to 12, 77 to 80 (black closed and open circles) correspond to transposases. Colors and vertical bars define the three main regions that contain duplicated genes. Duplications 1, 9 and 13 are not shown because both repeated copies map to the pAA plasmid (not shown in the figure). Duplications 4 to 8 contain one copy in the chromosome (shown) and the other in the plasmid (not shown). **b** Details of regions 1 to 3, showing duplicated genes (labeled in black). To show inversion, both copies of region 2 are shown. Genes labeled in green correspond to the *flu yeeR* gene cluster. Thin gray arrows correspond to the in silico operon prediction. The figure was generated using Easyfig [40]. See Additional file 1: Table S3 for the function of each duplicated gene

hha null and *hns* derivatives. This result was also observed for several genes in region 1. Only two genes from region 3 appear to be coregulated by H-NS/Hha.

Genes from 042 regions 1 and 2 are also duplicated in several other pathogenic *E. coli* strains

After determining the extent of gene duplications in the genome of strain 042, we addressed the question of whether the existing duplicates in strain 042 were strain-specific or whether they were generated in some putative ancestor and are also present in many other *E. coli* strains. We used the DNA sequences of the selected 28 *E. coli* genomes to perform a gene duplication analysis (see the methods for details) and annotated the number of copies of each of the duplicated genes from strain 042 that were detected in each of the other

Table 1 Comparative expression of the duplicated genes from regions 1 to 3 of strain 042 in the *hha* null (deletion of the *hha* and *hha2* alleles) and *hns* mutants. Values indicate the fold change with respect to the wt strain. Fold change values higher than two are considered significant

	group	locus tag	*hha null*	*hns*
Region 1	21	EC042_1328	2.1	6
	23	EC042_1330	2.6	5
	24	EC042_1333	2.5	4.2
	25	EC042_1336	1.8	0.6
	26	EC042_1342	3.5	4.8
	27	EC042_1343	2.8	4.8
	28	EC042_1344	1.7	4
	29	EC042_1349	2.4	3.8
	30	EC042_1353	4.3	3.3
	31	EC042_1371	1.5	6.1
	32	EC042_1372	1.5	3.8
	33	EC042_1373	1.6	3.3
	34	EC042_1376	2	3.7
	35	EC042_1377	2.4	4.2
Region 2	46	EC042_2236A	2.9	2.9
	47	EC042_2237	3.8	4.5
	48	EC042_2238	5.6	4.2
	49	EC042_2239	3.8	3
	50	EC042_2241	6.2	4.8
	51	EC042_2242	3.1	3.9
	52	EC042_2243	3	3.8
	53	EC042_2244	2.2	2.2
	54	EC042_2244A	4.9	5.9
	55	EC042_2245	5	5.1
	56	EC042_2246	4.6	4.8
	57	EC042_2247	5	3.9
	58	EC042_2247A	5	4.7
	59	EC042_2248	5.1	4.5
	60	EC042_2249	4.9	2.6
Region 3	64	EC042_3180	3	4.4
	65	EC042_3181	3.4	2.4
	66	EC042_3182	2.1	2.2
	67	EC042_3183	1.9	1.3
	68	EC042_3187	3.8	1.1
	70	EC042_3190	0.6	0.3
	71	EC042_3191	1.9	1.6

genomes (Fig. 4). With respect to region 1, 10 out of the 14 duplicated genes in strain 042 were duplicated in most of the genomes analyzed. With respect to region 2, the 15 duplicated genes are duplicated in either some or most of the genomes studied (Fig. 4). Six of the genes from that region (listed as 55 to 60), which appear as a single linkage group and belong to the same putative transcriptional unit, are present in several copies (4 to 6) in the genomes of most of the strains. These genes encode conserved hypothetical proteins (55, 58–60), a putative antirestriction protein (56) and a putative DNA repair protein (57). With regard to region 3, duplications of the eight genes identified in strain 042 are a specific feature of that strain. Several of these genes are either absent or present in a single copy in most of the genomes studied (Fig. 4). It is relevant to mention here that only 9 out of 40 duplicated genes from strain 042 that map to regions 1 to 3 are present in a single copy in the genome of strain MG1655. The rest of the genes are not present in the genome of the commensal strain.

Gene duplications in the genomes of strains CFT073 (UPEC) and O145:H28 (EHEC)

To obtain a more complete picture of gene duplications in *E. coli*, we decided to analyze the genomes of two other *E. coli* strains that belong to pathotypes different from that of strain 042. Strain CFT073 is uropathogenic (UPEC), and strain O145:H28 is enterohemorrhagic (EHEC). With respect to strain CFT073, 94 duplicated genes could be identified. They can be grouped into six different DNA regions (Fig. 5, Additional file 1: Table S4). Some of these genes correspond to transposases, similar to strain 042.

A total of 154 duplicated genes could be identified in the genome of strain O145:H28. The duplicated genes can also be grouped into six regions (Fig. 6, Additional file 1: Table S5). In this strain, several of the identified genes are present in more than two copies (Fig. 6). After identifying the duplicated genes in strains CFT073 and O145:H28, we also determined which of them are also duplicated in other *E. coli* strains. The genomic DNA sequences of the 28 *E. coli* strains were used to perform gene duplication analysis, and the number of copies of each of the duplicated genes from strains CFT073 and O145:H28 that were detected in each genome of the 28 *E. coli* strains was annotated. For both strains, duplicated genes that also occur as duplicates in other *E. coli* strains correspond to those already identified in strain 042 (Additional file 1: Figures S3 and S4). Some genes appear to be strain specific, as observed for strain 042. Some duplications in strains CFT073 and O145:H28 revealed a novel pattern: they are pathotype specific. The duplicated genes from strain CF703 region 5 belong to that group. Interestingly, most of these genes encode putative fimbrial proteins (Additional file 1: Table S4). Another example corresponds to duplications mapping in the region 4 of strain O145:H28. These genes are duplicated only in all the EHEC

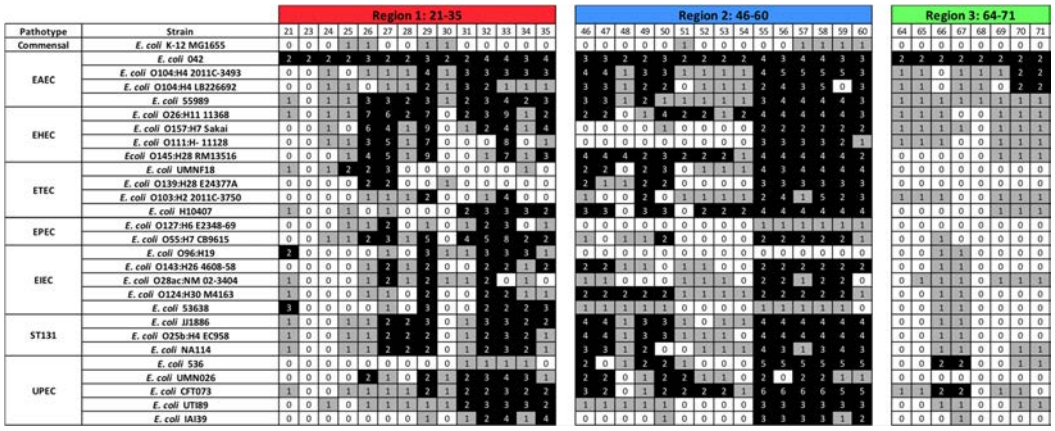


Fig. 4 Distribution of the detected duplicated genes in other *E. coli* strains belonging to a wide range of pathotypes. BLASTp analysis was used for the study. White color, gene absent. Gray color, gene present in a single copy. Black color, gene present in two or more copies. The numbers show the copy number of each gene. Note that a significant number of the duplicated genes are absent in strain MG1655

strains, one EPEC strain and one ETEC strain. Several of those genes are phage genes (Additional file 1: Table S5). In contrast to the EAEC strain 042, duplicated genes in the EHEC strain O145:H28 are not duplicated in the UPEC strains.

Discussion

The existence of gene duplications in both eukaryotes and prokaryotes has been extensively studied [18–23, 29, 30]. Several reports have established the basis for how gene duplication and divergence generate families and

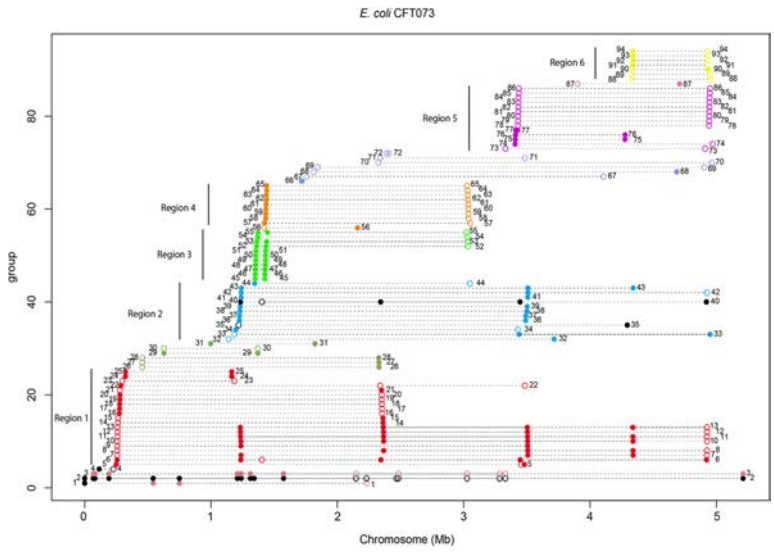


Fig. 5 Genes duplicated in the *E. coli* strain CFT073. The X axis corresponds to the linear map of the chromosome. Each group of spots connected by a horizontal dotted line corresponds to a single gene, which is shown to be duplicated in different positions of the chromosome. The different spots indicate the map positions of the different copies of the gene. Point shapes represent the strand on which a protein is encoded: filled circle for (+) strand and circle for (–) strand. Numbers correspond to the different duplicated genes, which have been numbered by their order starting from the origin of the chromosomal map. Black closed and open circles correspond to transposases. Colors and vertical bars define the six main regions that contain duplicated genes. See Additional file 1: Table S4 for the function of each of the duplicated genes

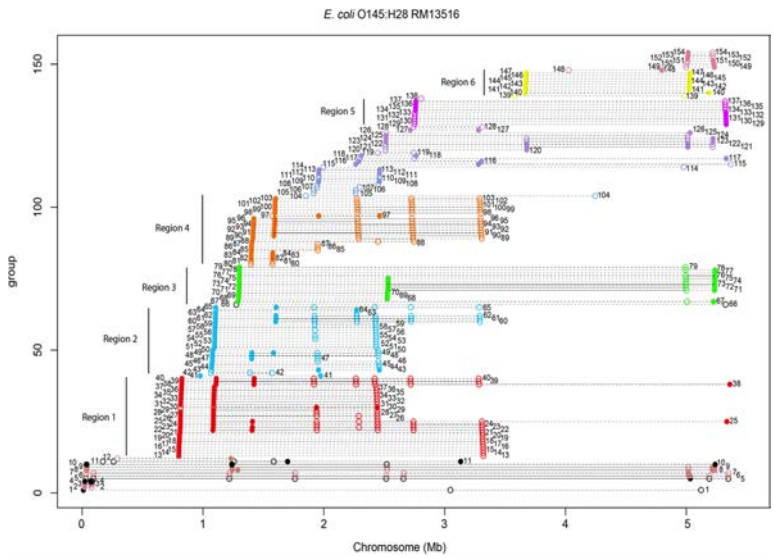


Fig. 6 Genes duplicated in the *E. coli* strain O145:H28. The X axis corresponds to the linear map of the chromosome. Each group of spots connected by a horizontal dotted line corresponds to a single gene, which is shown to be duplicated in different positions of the chromosome. The different spots indicate the map positions of the different copies of the gene. Point shapes represent the strand on which a protein is encoded: filled circle for (+) strand and circle for (−) strand. Numbers correspond to the different duplicated genes, which have been numbered by their order starting from the origin of the chromosomal map. Black closed and open circles correspond to transposases. Colors and vertical bars define the six main regions that contain duplicated genes. See Additional file 1: Table S5 for the function of each duplicated gene

superfamilies of proteins [21]. Gene duplications have been associated to the adaptation of cells to a changing environment [31, 32], and have been found to occur more frequently among HGT genes than among indigenous genes [33]. The presence of several copies of genes such as *flu* in some *E. coli* strains was previously reported [30, 32, 34–36].Nevertheless, detailed information about the extent of gene duplications in the genomes of the different types of pathogenic *E. coli* strains is needed. We applied an extensive blast search to identify putative internal duplications in the 042 strain using a moderate parameter cutoff (BLAST cutoff: > 85% similarity, > 85% alignment length and *e*-value $< 10^{-10}$) and found that most duplicates cluster together in specific regions of the 042 genome. The results obtained suggest that different mechanisms underlie these duplication events. Whereas the duplication of region 2 from strain 042 involves an inversion, this was not the case for regions 1 and 3. Interestingly, the genes in region 1 have a phage origin and are widespread in several strains. It is worth mentioning that, as a general rule, duplications result in the presence of copies of the duplicated gene in both strands of the *E. coli* chromosome. It is also remarkable that a significant number of the duplicated genes are organized in putative transcriptional units (Fig. 3b), thus suggesting the existence of coordinated expression in response to specific stimuli.

The comparative analysis of gene duplications in *E. coli* strains belonging to different pathotypes provides relevant information that can contribute to our understanding of the virulence mechanisms of this pathogen and better establish the relationships among the *E. coli* pathotypes. The existence of a significant number of genes that are duplicated in a wide range of pathotypes but absent from commensal strains suggests that these genes can play a relevant role in *E. coli* virulence. Genes 55 to 60 from strain 042 region 2 are duplicated in all except three of the 26 pathogenic *E. coli* strains analyzed. Given that detailed information about the function of the products encoded by a large number of these genes is missing, assigning functions to them and to many other genes of unknown function is a critical issue for better understanding the ability of *E. coli* to cause disease.

In addition to identifying a set of duplicated genes that is widespread in the different *E. coli* pathotypes, our study provides additional novel information on genomic features of virulent *E. coli* strains. In *E. coli,* some gene duplication processes are restricted to either specific strains or specific pathotypes. Examples are the duplicated genes in region 3

of strain 042 or the duplicated genes in regions 5 and 4 from the UPEC strain CFT073 and the EHEC strain 0145:H28, respectively. The study of the function of these genes can also contribute to a better understanding of the mechanisms underlying virulence in these pathotypes. It is well known that UPEC strains express specific types of fimbriae. Some of the duplicated genes in region 5 from the UPEC strain CFT073 encode putative fimbrial proteins, which might play a role in UPEC pathogenesis.

The correlation we observed between *hha* duplication and the presence of the duplicated *yeeR irmA* (*aec69*) gene cluster suggested that Hha (and H-NS) could modulate the expression of duplicated *E. coli* genes. The analysis of the comparative expression of duplicated genes in the wt 042 strain and its isogenic *hha* null and *hns* derivatives shows that under specific growth conditions (LB medium, 37 °C), H-NS/Hha proteins downregulate the expression of a significant number of duplicated genes. These data highlight a novel role for the H-NS/Hha proteins in silencing several of the genes that are duplicated in strain 042. Hence, it can be hypothesized that to avoid fitness costs, duplications of genes targeted by global regulators may require the duplication of the genes that encode them. Derepression of H-NS/Hha-silenced genes can occur when environmental conditions change. Then, gene duplication may be advantageous because the two copies can exhibit different expression patterns and/or respond to different stimuli. This is the case for the duplicated *irmA* gene in strain 042 (our unpublished results).

A relevant point is whether HGT processes are underlying the presence of gene duplications in strain 042. The duplicated genes that map in the region 1 of strain 042 are of phage origin and can hence be considered as HGT DNA. In any case, a detailed phylogenetic analysis is being undertaken now to assess the origin of all duplicates that map in the three regions identified in strain 042.

Finally, our study has also shown some novel relationships between *E. coli* pathotypes. It is remarkable that most of the duplicated genes in the EAEC strain 042 are also duplicated in UPEC strains. Previous studies have suggested a close relationship between EAEC and UPEC strains [33, 37]. In fact, *E. coli* strains showing a hybrid UPEC/EAEC genotype have been isolated [38]. The similar gene duplication patterns of EAEC and UPEC strains further support this EAEC/UPEC relationship. Unlike EAEC strain 042, duplicated genes in the EHEC strain O145:H28 are usually duplicated in EPEC and ETEC strains but not in UPEC strains. A distinctive feature of EHEC strains is that some of the duplicated genes are present in more than two copies.

For some *E. coli* infections, such as those caused by ETEC, the effectiveness of the existing vaccines must be significantly improved [39]. If any of the gene products encoded by the identified duplicated genes are antigenic, they could be candidates for developing novel improved *E. coli* vaccines.

Conclusions

Duplications of the *hha* gene can be correlated with the presence of genes belonging to the *flu yeeR aec* gene cluster, which is also duplicated in several pathogenic *E. coli* strains. The analysis of gene duplications in the *E. coli* genome has shown that (i) a number of duplicated genes are widely distributed among pathogenic *E. coli* strains, irrespective of the pathotype; (ii) some duplicated genes are only present in specific pathotypes; and (iii) some duplicated genes are strain specific. The present study also shows a relationship between duplications of both genes encoding regulators and genes encoding their targets. Our study also shows novel relationships between *E. coli* pathotypes. Finally, the distribution of duplicated genes in a high percentage of pathogenic *E. coli* isolates suggests that these genes must play a role in virulence. Hence, some of their gene products can serve as new targets for combating *E. coli* infections.

Additional files

Additional file 1: Table S1. List of *E. coli* strains whose genomes have been used. **Table S2.** Distribution of genes from the *flu yeeR irmA aec70 aec71*. **Table S3.** Locus tag and gene function of each of the duplicated genes in regions 1, 2 and 3 of strain 042. **Table S4.** Locus tag and gene function of each of the duplicated genes in regions 1, 2, 3, 4, 5 and 6 of strain CFT073. **Table S5.** Locus tag and gene function of each of the duplicated genes in regions 1, 2, 3, 4, 5 and 6 of strain O145:H8. The locus tags of the different copies are shown. **Figure S1.** Five-set Venn diagram of the exclusive core-genome of the *hha2*/3^{+} set (*E. coli* strains 042, NA114, O104:H4 LB226692, ETEC H10407 and UMN026). **Figure S2.** Genes duplicated in the *E. coli* strain 042, identified by using BLASTn instead of BLASTp. **Figure S3.** Distribution of the strain CFT073 duplicated genes in other *E. coli* strains belonging to a wide range of pathotypes. **Figure S4.** Distribution of strain O145:H28 duplicated genes in other *E. coli* strains belonging to a wide range of pathotypes. (DOC 2532 kb)

Additional file 2: DNA sequences of the genes comprising the three shared families identified in the exclusive core genome of the *hha2/hha3*$^{+}$ set (strains 042, NA114, O104:H4 2011C-3493, ETEC H10407 and UMN026). (DOCX 73 kb)

Abbreviations

BLAST: Basic Local Alignment Search Tool; *E. coli*: *Escherichia coli*; EAEC: Enteroaggregative *Escherichia coli*; EHEC: Enterohemorrhagic *Escherichia coli*; EIEC: Enteroinvasive *Escherichia coli*; EPEC: Enteropathogenic *Escherichia coli*; ETEC: Enterotoxigenic *Escherichia coli*; UPEC: Uropathogenic *Escherichia coli*

Acknowledgements

A.P. and J.F.S.H. were supported by Formación del Profesor Universitario (FPU) fellowships (Ministerio de Educación de España), and M.B. was supported by an FI fellowship from the Generalitat de Catalunya.

Funding
This work was supported by Ministerio de Economıa y Competitividad of Spain grants (CGL2016–75255 to JR, BIO2016–76412-C2–1-R (AEI/FEDER, UE) to AJ, and CERCA Programme/Generalitat de Catalunya to AJ).

Availability of data and materials
The RNA sequencing reads had been deposited in the Gene Expression Omnibus (GEO) Sequence Read Archive of the National Center for Biotechnology Information (GSE105133) under accession numbers GSM2822965, GSM2822966, and GSM2822967.

Authors' contributions
MB, JFS, AP and PH performed the experimental work. AJ, MH and JR conceived the experiments and wrote the manuscript. All co-authors read and reviewed the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Department of Genetics, Microbiology and Statistics, University of Barcelona, Barcelona, Spain. [2]Biodiversity Research Institute (IRBio), University of Barcelona, Barcelona, Spain. [3]Institute of Biotechnology and Biomedicine (IBB), Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain. [4]Institute for Bioengineering of Catalonia, The Barcelona Institute of Science and Technology, Barcelona, Spain.

Received: 29 October 2018 Accepted: 10 April 2019
Published online: 24 April 2019

References

1. Kaper JB, Nataro JP, Mobley HL. Pathogenic *Escherichia coli*. Nat Rev Microbiol. 2004;2:123–40.
2. Nataro JP, Kaper JB, Robins-Browne R, Prado V, Vial P, Levine MM. Patterns of adherence of diarrheagenic *Escherichia coli* to HEp-2 cells. Pediatr Infect Dis J. 1987;6:829–31.
3. Okeke IN, Wallace-Gadsden F, Simons HR, Matthews N, Labar AS, Hwang J, et al. Multi-locus sequence typing of enteroaggregative *Escherichia coli* isolates from Nigerian children uncovers multiple lineages. PLoS One. 2010; 5:e14093.
4. Frank C, Werber D, Cramer JP, Askar M, Faber M, an der Heiden M, et al. Epidemic profile of Shiga-toxin-producing *Escherichia coli* O104:H4 outbreak in Germany. N Engl J Med. 2011;365:1771–80.
5. Bielaszewska M, Mellmann A, Zhang W, Köck R, Fruth A, Bauwens A, et al. Characterisation of the *Escherichia coli* strain associated with an outbreak of haemolytic uraemic syndrome in Germany, 2011: a microbiological study. Lancet Infect Dis. 2011;11:671–6.
6. Mayer CL, Leibowitz CS, Kurosawa S, Stearns-Kurosawa DJ. Shiga toxins and the pathophysiology of hemolytic uremic syndrome in humans and animals. Toxins (Basel). 2012;4:1261–87.
7. Nataro JP, Kaper JB. Diarrheagenic *Escherichia coli*. Clin Microbiol Rev. 1998; 11:142–201.
8. Nataro JP, Deng Y, Cookson S, Cravioto A, Savarino SJ, Guers LD, et al. Heterogeneity of enteroaggregative *Escherichia coli* virulence demonstrated in volunteers. J. Infect. Dis. 1995;171:465–8.
9. Chaudhuri RR, Sebaihia M, Hobman JL, Webber MA, Leyton DL, Goldberg MD, et al. Complete genome sequence and comparative metabolic profiling of the prototypical enteroaggregative *Escherichia coli* strain 042. PLoS One. 2010;5:e8801.
10. Nataro JP, Scaletsky IC, Kaper JB, Levine MM, Trabulsi LR. Plasmid-mediated factors conferring diffuse and localized adherence of enteropathogenic *Escherichia coli*. Infect Immun. 1985;48:378–83.
11. Nataro JP, Yikang D, Yingkang D, Walker K. AggR, a transcriptional activator of aggregative adherence fimbria I expression in enteroaggregative *Escherichia coli*. J Bacteriol. 1994;176:4691–9.
12. Czeczulin JR, Balepur S, Hicks S, Phillips A, Hall R, Kothary MH, et al. Aggregative adherence fimbria II, a second fimbrial antigen mediating aggregative adherence in enteroaggregative *Escherichia coli*. Infect Immun. 1997;65:4135–45.
13. Morin N, Tirling C, Ivison SM, Kaur AP, Nataro JP, Steiner TS. Autoactivation of the AggR regulator of enteroaggregative *Escherichia coli* in vitro and in vivo. FEMS Immunol Med Microbiol. 2010;58:344–55.
14. Prieto A, Urcola I, Blanco J, Dahbi G, Muniesa M, Quirós P, et al. Tracking bacterial virulence: global modulators as indicators. Sci Rep. 2016;6:25973.
15. Madrid C, Balsalobre C, García J, Juárez A. The novel Hha/YmoA family of nucleoid-associated proteins: use of structural mimicry to modulate the activity of the H-NS family of proteins. Mol Microbiol. 2007;63:7–14.
16. Madrid C, García J, Pons M, Juárez A. Molecular evolution of the H-NS protein: interaction with Hha-like proteins is restricted to enterobacteriaceae. J Bacteriol. 2007;189:265–8.
17. Shintani M, Suzuki-Minakuchi C, Nojiri H. Nucleoid-associated proteins encoded on plasmids: occurrence and mode of function. Plasmid. 2015;80: 32–44.
18. Zhang J. Evolution by gene duplication: an update. Trends Ecol Evol. 2003; 18:292–8.
19. He X, Zhang J. Gene complexity and gene duplicability. Curr Biol. 2005;15: 1016–21.
20. Conant GC, Wolfe KH. Turning a hobby into a job: how duplicated genes find new functions. Nat. Rev. Genet. 2008;9:938–50.
21. Serres MH, Kerr ARW, McCormack TJ, Riley M. Evolution by leaps: gene duplication in bacteria. Biol Direct. 2009;4:46.
22. Innan H, Kondrashov F. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet. 2010;11:97–108.
23. Gao Y, Zhao H, Jin Y, Xu X, Han G-Z. Extent and evolution of gene duplication in DNA viruses. Virus Res. 2017;240:161–5.
24. Miele V, Penel S, Duret L. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics. 2011;12:116.
25. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
26. R Development Core Team (2008). R-project.org. Vienna, Austria; http://www.R-project.org
27. Prieto A, Bernabeu M, Aznar S, Ruiz-Cruz S, Bravo A, Queiroz MH, et al. Evolution of bacterial global modulators: role of a novel H-NS paralogue in the Enteroaggregative *Escherichia coli* strain 042. MSystems. 2018;3(3):e00220–17.
28. Moriel DG, Heras B, Paxman JJ, Lo AW, Tan L, Sullivan MJ, et al. Molecular and structural characterization of a novel *Escherichia coli* interleukin receptor mimic protein. MBio. 2016;7:e02046.
29. Walsh JB. How often do duplicated genes evolve new functions? Genetics. 1995;139:421–8.
30. Arun PVPS, Miryala SK, Chattopadhyay S, Thiyyagura K, Bawa P, Bhattacharjee M, et al. Identification and functional analysis of essential, conserved, housekeeping and duplicated genes. FEBS Lett. 2016;590: 1428–37.
31. Kondrashov FA. Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc Biol Sci. 2012;279:5048–57.
32. Elliott KT, Cuff LE, Neidle EL. Copy number change: evolving views on gene amplification. Future Microbiol. 2013;8:887–99.
33. Hooper SD, Berg OG. Duplication is more common among laterally transferred genes than among indigenous genes. Genome Biol. 2003;4:R48.
34. Restieri C, Garriss G, Locas M-C, Dozois CM. Autotransporter-encoding sequences are phylogenetically distributed among *Escherichia coli* clinical isolates and reference strains. Appl Environ Microbiol. 2007;73:1553–62.
35. Roche A, McFadden J, Owen P. Antigen 43, the major phase-variable protein of the *Escherichia coli* outer membrane, can exist as a family of proteins encoded by multiple alleles. Microbiology (Reading, Engl). 2001; 147:161–9.
36. van der Woude MW, Henderson IR. Regulation and function of Ag43 (*flu*). Annu Rev Microbiol. 2008;62:153–69.
37. Regua-Mangia AH, Irino K, da Silva Pacheco R, Pimentel Bezerra RM, Santos Périssé AR, Teixeira LM. Molecular characterization of uropathogenic and

diarrheagenic *Escherichia coli* pathotypes. J Basic Microbiol. 2010;50(Suppl 1): S107–15.
38. Lara FBM, Nery DR, de Oliveira PM, Araujo ML, Carvalho FRQ, Messias-Silva LCF, et al. Virulence markers and phylogenetic analysis of *Escherichia coli* strains with hybrid EAEC/UPEC genotypes recovered from sporadic cases of Extraintestinal infections. Front Microbiol. 2017;8:146.
39. Zhang W, Sack DA. Current Progress in developing subunit vaccines against Enterotoxigenic *Escherichia coli*-associated diarrhea. Clin Vaccine Immunol. 2015;22:983–91.
40. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. Bioinformatics. 2011;27:1009–10.

Additional File 1

Gene duplications in the *E. coli* genome: common themes among pathotypes

Bernabeu, M.[1], Sanchez-Herrero, J.F.[1,2], Huedo, P.[3], Prieto, A.[1], Hüttener, M.[1], Rozas, J.[1,2] and Juárez, A.[1,4*]

[1]Department of Genetics, Microbiology and Statistics, University of Barcelona, Barcelona, Spain.

[2]Biodiversity Research Institute (IRBio), University of Barcelona, Barcelona, Spain.

[3]Institute of Biotechnology and Biomedicine (IBB), Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain.

[4]Institute for Bioengineering of Catalonia, The Barcelona Institute of Science and Technology, Barcelona, Spain.

*Corresponding author: Prof. Antonio Juárez (ajuarez@ub.edu).

SUPPLEMENTARY TABLES

Table S1.

Pathotype	Strain	Presence of *hha2/3*	Accession number	Assembly Accesion
Commensal	*E. coli* K-12 MG1655	*hha2/3*$^-$	AJGD00000000.1	GCF_000482265.1
EAEC	*E.coli* 042	*hha2/3*$^+$	NC_017626.1	GCF_000027125.1
	E.coli O104:H4 2011C-3493	*hha2/3*$^+$	NC_018658.1	GCF_000299455.1
	E. coli O104:H4 LB226692	*hha2/3*$^+$	AFOB00000000.2	GCA_000215685.3
	E.coli 55989	*hha2/3*$^+$	NC_011748.1	GCF_000026245.1
EHEC	*E. coli* O26:H11 11368	*hha2/3*$^+$	NC_013361.1	GCF_000091005.1
	E. coli O157:H7 Sakai	*hha2/3*$^-$	NC_002695.1	GCF_000008865.1
	E. coli O111:H- 11128	*hha2/3*$^-$	AP010960.1	GCF_000010765.1
	E. coli O145:H28 RM13516	*hha2*$^+$	CP006262.1	GCF_000520055.1
ETEC	*E. coli* UMNF18	*hha2/3*$^-$	AGTD00000000.1	GCF_000220005.1
	E. coli O139:H28 E24377A	*hha3*$^+$	JXRF00000000.1	GCF_000017745.1
	E. coli O103:H2 2011C-3750	*hha2/3*$^-$	JHLL00000000.1	GCF_000616345.2
	E. coli H10407	*hha3*$^+$	FN649414.1	GCF_000210475.1
EPEC	*E. coli* O127:H6 E2348-69	*hha2/3*$^-$	NC_011601.1	GCF_000026545.1
	E. coli O55:H7 CB9615	*hha2*$^+$	CP001846.1	GCF_000025165.1
EIEC	*E. coli* O96:H19	*hha2*$^+$	JHNY01000124.1	GCF_001007915.1
	E. coli O143:H26 4608-58	*hha2/3*$^+$	JTCO01000000	GCF_000805835.1
	E. coli O28ac:NM 02-3404	*hha2/3*$^+$	JHNY00000000	GCF_000617165.2
	E. coli O124:H30 M4163	*hha2/3*$^+$	JTCN01000000	GCF_000805815.1
	E. coli 53638	*hha2/3*$^-$	AAKB00000000.2	GCF_000167915.2
ST131	*E. coli* JJ1886	*hha2/3*$^+$	CP006784.1	GCF_000493755.1
	E. coli O25b:H4 EC958	*hha2/3*$^-$	HG941718.1	GCF_000285655.3
	E. coli NA114	*hha2/3*$^+$	MIPU00000000.1	GCF_000214765.2
UPEC	*E. coli* 536	*hha2/3*$^-$	CP000247.1	GCF_000013305.1
	E. coli UMN026	*hha2*$^+$	CU928163.2	GCF_000026325.1
	E. coli CFT073	*hha2/3*$^+$	AE014075.1	GCF_000007445.1
	E. coli UTI89	*hha3*$^+$	CP000243.1	GCF_000013265.1
	E. coli IAI39	*hha2/3*$^-$	NC_011750.1	GCF_000026345.1

Table S1. List of *E. coli* strains whose genomes have been used for the different genomic analysis performed in this work. Pathotype, accession number and GenBank assembly accession details are indicated.

Table S2.

Pathotype	strain/gene	*flu*	*yeeR*	*irmA*	*aec70*	*aec71*	*hha2/3*
EAEC	*E. coli* 042	X	X	X	X	X	*hha*2/3$^+$
	E. coli O104_H4 2011C-3493	X	X	X	X	X	*hha*2/3$^+$
	E. coli O104:H4 LB226692		X	X	X	X	*hha*2/3$^+$
	E. coli 55989	X	X	X	X	X	*hha*2/3$^+$
EHEC	*E. coli* O26_H11_11368	X	X	X	X	X	*hha2/*3$^+$
	E. coli O145_H28 RM13516	X	X	X	X	X	*hha*2$^+$
ETEC	*E. coli* O139_H28_E24377A					X	*hha*3$^+$
	E. coli ETEC_H10407		X	X	X	X	*hha*3$^+$
EPEC	*E. coli* O55_H7_CB9615					X	*hha*2$^+$
EIEC	*E. coli* O96-H19						*hha*2$^+$
	E. coli O143-H26_4608-58	X	X			X	*hha2/*3$^+$
	E. coli O28ac-NM_02-3404	X	X	X		X	*hha2/*3$^+$
	E. coli O124-H30_M4163	X	X	X	X	X	*hha2/*3$^+$
ST131	*E. coli* JJ1886	X		X	X	X	*hha2/3*$^+$
	E. coli NA114		X	X	X	X	*hha*2/3$^+$
UPEC	*E. coli* UMN026	X	X	X		X	*hha*2$^+$
	E. coli CFT073	X	X	X	X	X	*hha*2/3$^+$
	E. coli UTI89					X	*hha*3$^+$
Commensal	*E. coli* K-12_MG1655	X					*hha*2/3$^-$
EHEC	*E. coli* O157_H7_Sakai					X	*hha*2/3$^-$
	E. coli O111_H-11128					X	*hha*2/3$^-$
ETEC	*E. coli* UMNF18		X	X	X	X	*hha*2/3$^-$
	E. coli O103-H2_2011C-3750	X	X	X	X	X	*hha*2/3$^-$
EPEC	*E. coli* O127_H6_E2348-69					X	*hha*2/3$^-$
EIEC	*E. coli* 53638					X	*hha*2/3$^-$
ST131	*E. coli* EC958	X	X	X			*hha*2/3$^-$
UPEC	*E. coli* 536	X				X	*hha*2/3$^-$
	E. coli IAI39					X	*hha*2/3$^-$

Table S2. Distribution of genes from the *flu yeeR irmA aec70 aec71* gene cluster among the 28 *E. coli* strains studied in this work.

	Group	Locus tag 1	Locus tag 2	Locus tag 3	Locus tag 4	Description
Region 1: 21-35	21	EC042_1328	EC042_2193			putative phage protein
	22	EC042_1329	EC042_2192			putative phage protein
	23	EC042_1330	EC042_2191			putative phage protein
	24	EC042_1333	EC042_2189			phage protein
	25	EC042_1336	EC042_1705	EC042_2186		putative host cell_killing modulation protein
	26	EC042_1342	EC042_2183			phage protein
	27	EC042_1343	EC042_2182			putative phage endodeoxyribonuclease
	28	EC042_1344	EC042_2181			phage antitermination protein
	29	EC042_1349	EC042_1703E	EC042_2175		putative phage lysozyme
	30	EC042_1353	EC042_1702			putative phage protein
	31	EC042_1371	EC042_1685			phage minor tail protein
	32	EC042_1372	EC042_1509	EC042_1684	EC042_2138	phage minor tail protein
	33	EC042_1373	EC042_1510	EC042_1683	EC042_2137	phage tail assembly protein
	34	EC042_1376	EC042_1512	EC042_2135		phage host specificity protein
	35	EC042_1377	EC042_1513	EC042_1680	EC042_2134	putative prophage_encoded outer membrane protein
Region 2: 46-60	46	EC042_2236A	EC042_4519	EC042_4793		conserved hypothetical protein
	47	EC042_2237	EC042_4518	EC042_4794		conserved hypothetical protein
	48	EC042_2238	EC042_4795			hypothetical protein
	49	EC042_2239	EC042_4798			conserved hypothetical protein
	50	EC042_2241	EC042_4512	EC042_4802		putative GTP_binding protein
	51	EC042_2242	EC042_4511			antigen 43 precursor (fluffing protein) (autotransporter)
	52	EC042_2243	EC042_4510			putative membrane protein
	53	EC042_2244	EC042_4509			putative exported protein
	54	EC042_2244A	EC042_4508A			conserved hypothetical protein
	55	EC042_2245	EC042_3221	EC042_4507	EC042_4805	conserved hypothetical protein
	56	EC042_2246	EC042_3222	EC042_4506		putative antirestriction protein
	57	EC042_2247	EC042_3223	EC042_4505	EC042_4807	putative DNA repair protein
	58	EC042_2247A	EC042_3224	EC042_4504	EC042_4808	conserved hypothetical protein
	59	EC042_2248	EC042_3225	EC042_4503		conserved hypothetical protein
	60	EC042_2249	EC042_3226	EC042_4502		conserved hypothetical protein
Region 3: 64-71	64	EC042_3180	EC042_4556			conserved hypothetical protein
	65	EC042_3181	EC042_4555			putative transcriptional regulator
	66	EC042_3182	EC042_4554			ParB_like nuclease
	67	EC042_3183	EC042_4553			conserved hypothetical protein
	68	EC042_3187	EC042_4548			putative helicase
	69	EC042_3189	EC042_4523			phage protein
	70	EC042_3190	EC042_4522			conserved hypothetical protein
	71	EC042_3191	EC042_4521			putative DNA_binding protein

Table S3. Locus tag and gene function of each of the duplicated genes in regions 1, 2 and 3 of strain 042. The locus tags of the different copies are shown. Gene functions correspond to locus tag 1.

Table S4. Locus tag and gene function of each of the duplicated genes in regions 1, 2, 3, 4, 5 and 6 of strain CFT073. The locus tags of the different copies are shown. Gene functions correspond to locus tag 1.

	Group	Locus tag 1	Locus tag 2	Locus tag 3	Locus tag 4	Locus tag 5	Locus tag 6	Gene function
Region 1: 5-25	5	C_RS01210	C_RS17180	C_RS17295				IS66 family insertion sequence hypothetical protein
	6	C_RS01265	C_RS06025	C_RS07130	C_RS11895	C_RS17135	C_RS24345	ATPase AAA
	7	C_RS01270	C_RS06050	C_RS17440	C_RS21565	C_RS24365		toxin
	8	C_RS01275	C_RS12005	C_RS17435	C_RS21560	C_RS24370		type IV toxin_antitoxin system YeeU family antitoxin
	9	C_RS01280	C_RS06035	C_RS28815				hypothetical protein
	10	C_RS01285	C_RS06030	C_RS12000	C_RS17425	C_RS21555	C_RS24375	DUF987 domain_containing protein
	11	C_RS01290	C_RS06015	C_RS11995	C_RS17420	C_RS21550	C_RS24380	hypothetical protein
	12	C_RS01295	C_RS06010	C_RS11990	C_RS17415	C_RS24385		antirestriction protein
	13	C_RS01305	C_RS06005	C_RS11985	C_RS17410	C_RS21540	C_RS24390	DUF945 domain_containing protein
	14	C_RS01335	C_RS11970					phospholipase
	15	C_RS01340	C_RS11965					chemotaxis protein
	16	C_RS01380	C_RS11950					TonB_dependent siderophore receptor
	17	C_RS01385	C_RS11945					ABC transporter substrate_binding protein
	18	C_RS01390	C_RS11940					iron ABC transporter permease
	19	C_RS01395	C_RS11935					ABC transporter ATP_binding protein
	20	C_RS01400	C_RS11930					hypothetical protein
	21	C_RS26150	C_RS27170					hypothetical protein
	22	C_RS01440	C_RS11850	C_RS17315				hypothetical protein
	23	C_RS01505	C_RS26585					hypothetical protein
	24	C_RS01585	C_RS05670					hypothetical protein
	25	C_RS01595	C_RS05680					hypothetical protein
Region 2: 32-44	32	C_RS26480	C_RS27670					hypothetical protein
	33	C_RS05805	C_RS17055	C_RS24530				Major pilus subunit operon regulatory protein
	34	C_RS05855	C_RS16995					transcriptional regulator
	35	C_RS05950	C_RS21295					transposase
	36	C_RS05980	C_RS17355					50S ribosome_binding GTPase
	37	C_RS05985	C_RS17360					autotransporter domain_containing protein
	38	C_RS05990	C_RS17395					hypothetical protein
	39	C_RS05995	C_RS17400					hypothetical protein
	40	C_RS06020	C_RS07135	C_RS11890	C_RS17130	C_RS24340		IS21 family transposase ISEc10
	41	C_RS06055	C_RS17445					hypothetical protein
	42	C_RS06060	C_RS17450	C_RS24355				DUF957 domain_containing protein
	43	C_RS06070	C_RS17455	C_RS21580				DUF4942 domain_containing protein
	44	C_RS06735	C_RS15130					lysozyme
Region 3: 45-55	45	C_RS06780	C_RS07355					phage terminase large subunit family protein
	46	C_RS06785	C_RS07360					head_tail joining protein
	47	C_RS06790	C_RS07365					phage portal protein
	48	C_RS06795	C_RS07370					S49 family peptidase
	49	C_RS06800	C_RS07375					head decoration protein
	50	C_RS06805	C_RS07380					minor capsid protein E
	51	C_RS06815	C_RS07390					tail attachment protein
	52	C_RS06845	C_RS07420	C_RS15020				phage tail tape measure protein
	53	C_RS06860	C_RS07435	C_RS15005				phage tail protein
	54	C_RS06880	C_RS14990					Ail/Lom family protein
	55	C_RS06895	C_RS07460	C_RS14965				hypothetical protein

	Group	Locus tag 1	Locus tag 2	Gene function
Region 4: 56-65	56	C_RS07310	C_RS11125	phosphoporin PhoE
	57	C_RS07315	C_RS15140	lysis protein S
	58	C_RS07395	C_RS15045	tail protein
	59	C_RS07400	C_RS15040	tail protein
	60	C_RS07410	C_RS15030	phage minor tail protein G
	61	C_RS07415	C_RS15025	phage tail assembly protein T
	62	C_RS07425	C_RS15015	tail protein
	63	C_RS07430	C_RS15010	phage minor tail protein L
	64	C_RS07440	C_RS15000	tail assembly protein
	65	C_RS07445	C_RS14995	host specificity protein J
Region 5: 73-86	73	C_RS16485	C_RS24305	lysine__tRNA ligase
	74	C_RS16885	C_RS24645	DUF4102 domain_containing protein
	75	C_RS16890	C_RS21210	hypothetical protein
	76	C_RS27505	C_RS27835	hypothetical protein
	77	C_RS16905	C_RS16980	IS66 family insertion sequence hypothetical protein
	78	C_RS17000	C_RS24480	fimbrial protein
	79	C_RS17005	C_RS24485	fimbrial protein
	80	C_RS17010	C_RS24490	fimbrial protein
	81	C_RS17015	C_RS24495	type 1 fimbrial protein
	82	C_RS17020	C_RS24500	protein papJ
	83	C_RS17025	C_RS24505	molecular chaperone
	84	C_RS17030	C_RS24510	PapC/FimD family outer membrane usher protein
	85	C_RS17035	C_RS24515	type 1 fimbrial protein
	86	C_RS17045	C_RS24525	major pilus subunit operon regulatory protein PapB
Region 6: 88-94	88	C_RS21495	C_RS24540	AlpA family transcriptional regulator
	89	C_RS21500	C_RS24545	hypothetical protein
	90	C_RS21505	C_RS24445	inovirus Gp2 family protein
	91	C_RS21515	C_RS24410	GTPase
	92	C_RS21525	C_RS24400	WYL domain_containing protein
	93	C_RS21530	C_RS24395	hypothetical protein
	94	C_RS21535	C_RS28070	DUF905 domain_containing protein

Table S5. Locus tag and gene function of each of the duplicated genes in regions 1, 2, 3, 4, 5 and 6 of strain O145:H8. The locus tags of the different copies are shown. Gene functions correspond to locus tag 1.

	Group	Locus tag 1	Locus tag 2	Locus tag 3	Locus tag 4	Gene function
Region 1: 13-40	13	ECRM13516_RS03805	ECRM13516_RS16985			Replication protein P
	14	ECRM13516_RS03810	ECRM13516_RS16980			protein ren
	15	ECRM13516_RS03815	ECRM13516_RS16975			recombination protein NinB
	16	ECRM13516_RS03820	ECRM13516_RS16970			phage N_6_adenine_methyltransferase
	17	ECRM13516_RS03825	ECRM13516_RS30035			NinE family protein
	18	ECRM13516_RS03835	ECRM13516_RS16955			endonuclease
	19	ECRM13516_RS03840	ECRM13516_RS16950			protein ninG
	20	ECRM13516_RS03845	ECRM13516_RS16945			serine/threonine protein phosphatase
	21	ECRM13516_RS03875	ECRM13516_RS16935			hypothetical protein
	22	ECRM13516_RS03885	ECRM13516_RS07050	ECRM13516_RS14160	ECRM13516_RS16925	DUF1737 domain_containing protein
	23	ECRM13516_RS03890	ECRM13516_RS07060	ECRM13516_RS14155	ECRM13516_RS16920	holin
	24	ECRM13516_RS03895	ECRM13516_RS09815	ECRM13516_RS14150	ECRM13516_RS16915	DUF1327 domain_containing protein
	25	ECRM13516_RS03900	ECRM13516_RS07075	ECRM13516_RS14145	ECRM13516_RS16910	lysozyme
	26	ECRM13516_RS03905	ECRM13516_RS05270	ECRM13516_RS12555		hypothetical protein
	27	ECRM13516_RS03910	ECRM13516_RS05275	ECRM13516_RS09800	ECRM13516_RS11655	lysis protein
	28	ECRM13516_RS03915	ECRM13516_RS29045	ECRM13516_RS29455	ECRM13516_RS29675	hypothetical protein
	29	ECRM13516_RS03920	ECRM13516_RS05280	ECRM13516_RS09795	ECRM13516_RS12545	PerC family transcriptional regulator
	30	ECRM13516_RS03925	ECRM13516_RS29050	ECRM13516_RS29450	ECRM13516_RS29670	DUF3950 domain_containing protein
	31	ECRM13516_RS03935	ECRM13516_RS05285	ECRM13516_RS12540		terminase
	32	ECRM13516_RS03940	ECRM13516_RS05290	ECRM13516_RS12535		phage terminase large subunit family protein
	33	ECRM13516_RS03945	ECRM13516_RS05295	ECRM13516_RS12530		head_tail joining protein
	34	ECRM13516_RS03950	ECRM13516_RS05300	ECRM13516_RS12525		phage portal protein
	35	ECRM13516_RS03955	ECRM13516_RS05305	ECRM13516_RS12520		S49 family peptidase
	36	ECRM13516_RS03960	ECRM13516_RS05310	ECRM13516_RS12515		head decoration protein
	37	ECRM13516_RS03965	ECRM13516_RS05315	ECRM13516_RS12510		minor capsid protein E
	38	ECRM13516_RS03980	ECRM13516_RS09685	ECRM13516_RS12440	ECRM13516_RS14005	hypothetical protein
	39	ECRM13516_RS03985	ECRM13516_RS09680	ECRM13516_RS12435	ECRM13516_RS14000	tail fiber protein
	40	ECRM13516_RS03990	ECRM13516_RS09675	ECRM13516_RS12430	ECRM13516_RS13995	phage tail protein
Region 2: 41-65	41	ECRM13516_RS04700	ECRM13516_RS10020			dimethyl sulfoxide reductase subunit B
	42	ECRM13516_RS05115	ECRM13516_RS06915	ECRM13516_RS07975		exonuclease
	43	ECRM13516_RS05135	ECRM13516_RS09930	ECRM13516_RS12695		DUF1391 domain_containing protein
	44	ECRM13516_RS05160	ECRM13516_RS12680			Rha family transcriptional regulator
	45	ECRM13516_RS05165	ECRM13516_RS12675			hypothetical protein
	46	ECRM13516_RS05190	ECRM13516_RS09865	ECRM13516_RS12635		Hok/Gef family protein
	47	ECRM13516_RS31170	ECRM13516_RS31245	ECRM13516_RS31350		hypothetical protein
	48	ECRM13516_RS05200	ECRM13516_RS07025	ECRM13516_RS09855	ECRM13516_RS12620	DUF968 domain_containing protein
	49	ECRM13516_RS05205	ECRM13516_RS07030	ECRM13516_RS09850	ECRM13516_RS12615	endodeoxyribonuclease
	50	ECRM13516_RS05225	ECRM13516_RS12600			transcriptional regulator AraC
	51	ECRM13516_RS05320	ECRM13516_RS12505			hypothetical protein
	52	ECRM13516_RS05325	ECRM13516_RS12500			tail attachment protein
	53	ECRM13516_RS05330	ECRM13516_RS12495			tail protein
	54	ECRM13516_RS05335	ECRM13516_RS09735	ECRM13516_RS12490		tail protein
	55	ECRM13516_RS05340	ECRM13516_RS09730	ECRM13516_RS12485		tail protein
	56	ECRM13516_RS05345	ECRM13516_RS12480			phage minor tail protein G
	57	ECRM13516_RS05350	ECRM13516_RS09720	ECRM13516_RS12475		phage tail assembly protein T
	58	ECRM13516_RS05355	ECRM13516_RS12470			phage tail tape measure protein
	59	ECRM13516_RS05360	ECRM13516_RS09710	ECRM13516_RS12465		tail protein
	60	ECRM13516_RS05365	ECRM13516_RS09705	ECRM13516_RS12460	ECRM13516_RS14025	phage minor tail protein L
	61	ECRM13516_RS05370	ECRM13516_RS09700	ECRM13516_RS12455	ECRM13516_RS14020	phage tail protein
	62	ECRM13516_RS05375	ECRM13516_RS09695	ECRM13516_RS12450	ECRM13516_RS14015	tail assembly protein
	63	ECRM13516_RS05380	ECRM13516_RS29570			hypothetical protein
	64	ECRM13516_RS05385	ECRM13516_RS11550			superoxide dismutase
	65	ECRM13516_RS05390	ECRM13516_RS09690	ECRM13516_RS12445	ECRM13516_RS14010	host specificity protein J

	Group	Locus tag 1	Locus tag 2	Locus tag 3	Locus tag 4	Gene function
Region 3: 67-79	67	ECRM13516_RS06345	ECRM13516_RS25305	ECRM13516_RS26370		50S ribosome_binding GTPase
	68	ECRM13516_RS06355	ECRM13516_RS13030			hypothetical protein
	69	ECRM13516_RS06360	ECRM13516_RS13035			chemotaxis protein
	70	ECRM13516_RS06365	ECRM13516_RS13040			phospholipase
	71	ECRM13516_RS06375	ECRM13516_RS13055	ECRM13516_RS25280	ECRM13516_RS26395	DUF945 domain_containing protein
	72	ECRM13516_RS06380	ECRM13516_RS13060	ECRM13516_RS25270	ECRM13516_RS26400	antirestriction protein
	73	ECRM13516_RS06385	ECRM13516_RS13065	ECRM13516_RS25265	ECRM13516_RS26405	hypothetical protein
	74	ECRM13516_RS14095	ECRM13516_RS13070	ECRM13516_RS25260	ECRM13516_RS26410	DUF987 domain_containing protein
	75	ECRM13516_RS06400	ECRM13516_RS13075	ECRM13516_RS25255	ECRM13516_RS26420	type IV toxin_antitoxin system YeeU family antitoxin
	76	ECRM13516_RS06405	ECRM13516_RS25250	ECRM13516_RS26425		toxin
	77	ECRM13516_RS06410	ECRM13516_RS26430			hypothetical protein
	78	ECRM13516_RS06415	ECRM13516_RS25240	ECRM13516_RS26435		DUF957 domain_containing protein
	79	ECRM13516_RS06420	ECRM13516_RS25235			DUF4942 domain_containing protein
Region 4: 80-103	80	ECRM13516_RS06950	ECRM13516_RS07995			hypothetical protein
	81	ECRM13516_RS06960	ECRM13516_RS08005			transcriptional regulator
	82	ECRM13516_RS06965	ECRM13516_RS08010			transcriptional regulator
	83	ECRM13516_RS06970	ECRM13516_RS08015			hypothetical protein
	84	ECRM13516_RS31235	ECRM13516_RS31280			replication protein
	85	ECRM13516_RS06990	ECRM13516_RS09880			hypothetical protein
	86	ECRM13516_RS06995	ECRM13516_RS09875			hypothetical protein
	87	ECRM13516_RS31240	ECRM13516_RS31355			hypothetical protein
	88	ECRM13516_RS07045	ECRM13516_RS12580	ECRM13516_RS14170		membrane protein
	89	ECRM13516_RS07115	ECRM13516_RS11645	ECRM13516_RS14105	ECRM13516_RS16885	HNH endonuclease
	90	ECRM13516_RS07120	ECRM13516_RS11640	ECRM13516_RS14100	ECRM13516_RS16880	terminase small subunit
	91	ECRM13516_RS07125	ECRM13516_RS11635	ECRM13516_RS14095	ECRM13516_RS16875	terminase large subunit
	92	ECRM13516_RS07130	ECRM13516_RS11630	ECRM13516_RS14090	ECRM13516_RS16870	phage major capsid protein
	93	ECRM13516_RS07135	ECRM13516_RS11625	ECRM13516_RS14085	ECRM13516_RS16865	hypothetical protein
	94	ECRM13516_RS07145	ECRM13516_RS11610	ECRM13516_RS14070	ECRM13516_RS16850	phage gp6_like head_tail connector protein
	95	ECRM13516_RS07150	ECRM13516_RS11605	ECRM13516_RS14065	ECRM13516_RS16845	head_tail adaptor protein
	96	ECRM13516_RS07155	ECRM13516_RS11600	ECRM13516_RS14060	ECRM13516_RS16840	hypothetical protein
	97	ECRM13516_RS07980	ECRM13516_RS09950	ECRM13516_RS12715		cell division inhibitor
	98	ECRM13516_RS08140	ECRM13516_RS11595	ECRM13516_RS14055	ECRM13516_RS16835	DUF3168 domain_containing protein
	99	ECRM13516_RS08145	ECRM13516_RS11590	ECRM13516_RS14050	ECRM13516_RS16830	tail protein
	100	ECRM13516_RS08150	ECRM13516_RS11585	ECRM13516_RS14045	ECRM13516_RS16825	tail assembly chaperon
	101	ECRM13516_RS08155	ECRM13516_RS11580	ECRM13516_RS14040	ECRM13516_RS16820	phage tail protein
	102	ECRM13516_RS08160	ECRM13516_RS11575	ECRM13516_RS14035	ECRM13516_RS16815	phage tail tape measure protein
	103	ECRM13516_RS08165	ECRM13516_RS11570	ECRM13516_RS14030	ECRM13516_RS16810	tail protein
Region 5: 129-137	129	ECRM13516_RS14240	ECRM13516_RS26985			hypothetical protein
	130	ECRM13516_RS14245	ECRM13516_RS26980			RusA family crossover junction endodeoxyribonuclease
	131	ECRM13516_RS14250	ECRM13516_RS26975			LexA family transcriptional regulator
	132	ECRM13516_RS14265	ECRM13516_RS26965			helix_turn_helix domain_containing protein
	133	ECRM13516_RS29885	ECRM13516_RS30820			DUF4222 domain_containing protein
	134	ECRM13516_RS14295	ECRM13516_RS26915			hypothetical protein
	135	ECRM13516_RS14300	ECRM13516_RS26910			DUF2303 domain_containing protein
	136	ECRM13516_RS14305	ECRM13516_RS26905			HD family hydrolase
	137	ECRM13516_RS14310	ECRM13516_RS26900			hypothetical protein
Region 6: 139-147	139	ECRM13516_RS18210	ECRM13516_RS25205			lysine__tRNA ligase
	140	ECRM13516_RS18645	ECRM13516_RS26195			DUF4102 domain_containing protein
	141	ECRM13516_RS18665	ECRM13516_RS25425			RNA polymerase sigma factor FecI
	142	ECRM13516_RS18670	ECRM13516_RS25430			protein FecR
	143	ECRM13516_RS18675	ECRM13516_RS25435			TonB_dependent siderophore receptor
	144	ECRM13516_RS18680	ECRM13516_RS25440			Fe(3+)_dicitrate ABC transporter substrate_binding protein FecB
	145	ECRM13516_RS18685	ECRM13516_RS25445			Fe3+ dicitrate ABC transporter permease FecC
	146	ECRM13516_RS18690	ECRM13516_RS25450			Fe3+ dicitrate ABC transporter permease FecD
	147	ECRM13516_RS18695	ECRM13516_RS25455			Fe(3+) dicitrate transport ATP_binding protein FecE

SUPPLEMENTARY FIGURES

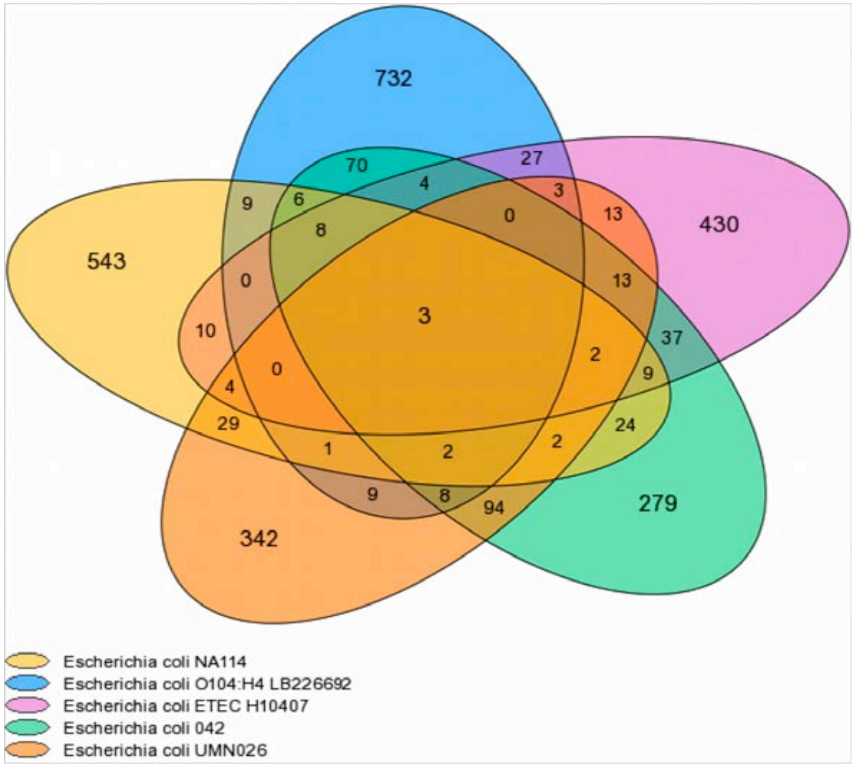


Figure S1. Five-set Venn diagram of the exclusive core-genome of the *hha*2/3^{+} set (*E. coli* strains 042, NA114, O104:H4 LB226692, ETEC H10407 and UMN026). Each ellipse displays the total number of coding sequences corresponding to a gene family of each *hha2/3*$^{+}$ strain which is not present in any genome of the five *hha2/3*$^{-}$ set (*E. coli* strains O111:H- 11128, 53638, IAI39, O127:H6 E2348/69 and O157:H7 Sakai). Intersections indicate shared genes between strains in a set. As shown, only three gene families are common to the five-strain set. No gene family of the *hha2/3*$^{-}$ set was exclusive and was present in the *hha2/3*$^{+}$ set (not shown in the figure).

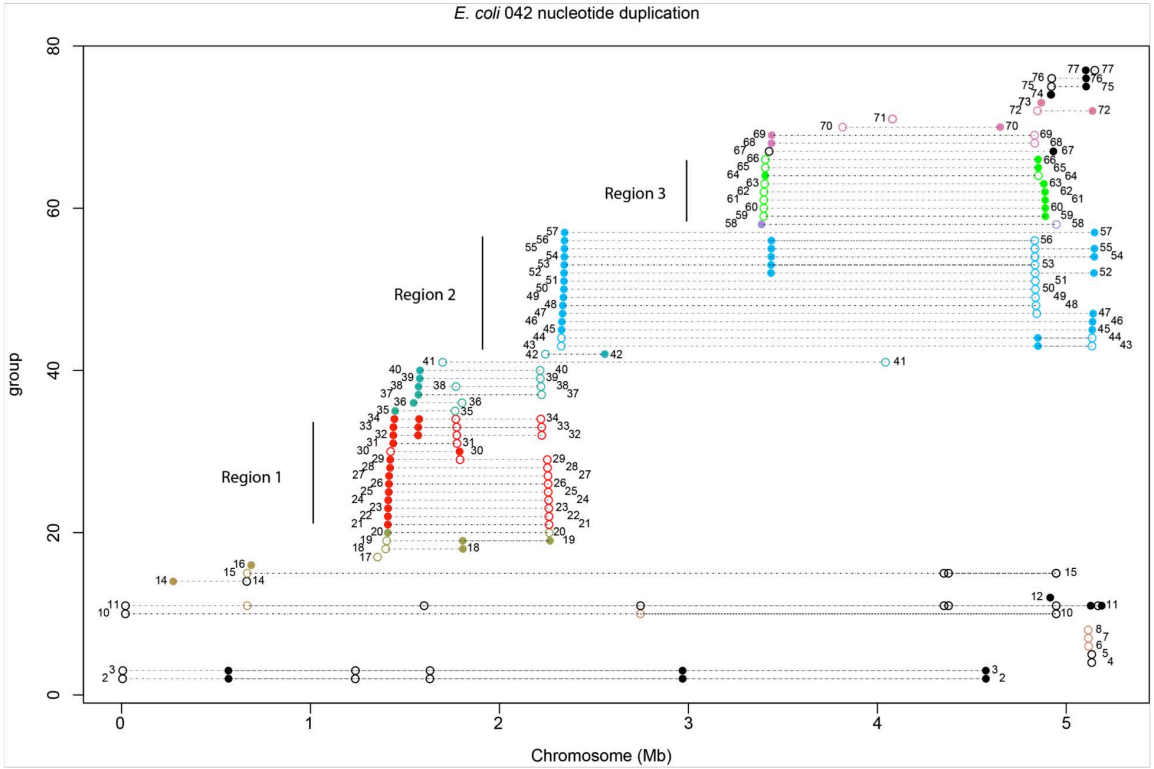


Figure S2. Genes duplicated in the *E. coli* strain 042, identified by using BLASTn instead of BLASTp. The X axis corresponds to the lineal map of the chromosome. Each group of spots connected by a horizontal dashed line corresponds to a single gene duplicated or amplified in different positions of the chromosome. The different spots indicate the map position of the different copies of the gene. Point shapes represent the strand on which a protein is codified: filled circle for (+) strand and circle for (-) strand. Numbers correspond to the different duplicated genes, which have been numbered by their order starting from the origin of the chromosomal map. Black closed and open circles correspond to transposases. Colors and vertical bars define the three main regions that contain duplicated genes.

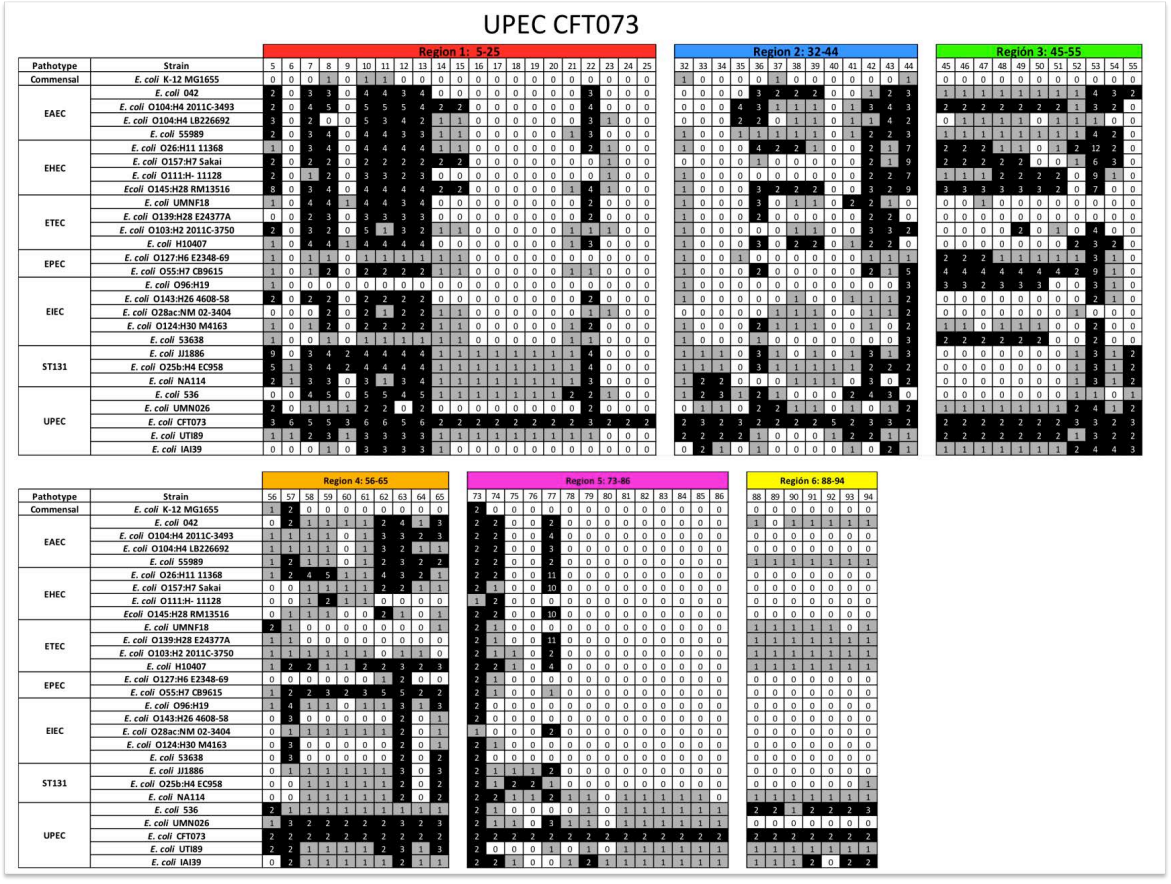


Figure S3. Distribution of the strain CFT073 duplicated genes in other E. coli strains belonging to a wide range of pathotypes. White color, gene absent. Grey color, gene present in a single copy. Black color, gene amplified. The numbers show the extent of gene duplication. Colours correspond to the identified regions.

Figure S4. Distribution of strain O145:H28 duplicated genes in other *E. coli* strains belonging to a wide range of pathotypes. White color, gene absent. Grey colur, gene present in a single copy. Black color, gene amplified. The numbers show the extent of gene amplification. Colors correspond to the identified regions.

Apéndice B

Financiación

Para el desarrollo de esta tesis doctoral el estudiante ha disfrutado de una beca de Formación del Profesorado Universitario (FPU) concedida por el Ministerio de Educación, Cultura y Deporte (MECD). El número de referencia es *FPU 2013-02607* y fue concedida por Resolución de 28 de noviembre de 2014, (BOE-8-12-2014) de la Secretaría de Estado de Educación, Formación Profesional y Universidades complementaria a la resolución de 22 de agosto de 2014 por renuncias de beneficiarios.

Por otra parte, la investigación llevada a cabo en la presente tesis ha estado financiada por diferentes proyectos:

1. Proyectos de ámbito nacional dentro del plan **Retos y Excelencia**:
   - **CGL2013-45211-C2-1-P**: *"Comprendiendo la base molecular de las radiaciones en islas: estudio evolutivo del sistema quimiosensorial en las arañas Dysdera de Canarias mediante transcriptómica"*. Investigador principal: Dr. Julio Rozas Liras. (Universitat de Barcelona). 2014-2016. Organismo responsable: Ministerio de Empresa, Industria y Competitividad (MINECO).
   - **CGL2016-75255-C2-2-P**: *"Genómica comparada y de la adaptación en el estudio de la radiación en Islas Macaronésicas: las arañas Dysdera en Canarias y su sistema quimiosensorial como sistema modelo"*. Investigador principal: Dr. Julio Rozas Liras. (Universitat de Barcelona). 2016-2019. Organismo responsable: MINECO.

2. Proyectos de ámbito nacional dentro de las acciones de dinamización **Redes de Excelencia**:

   - **CGL2015-71726-REDT**: *"Red Temática en Genómica de la Adaptación: AdaptNet"*. Investigador principal: Dr. Julio Rozas Liras. (Universitat de Barcelona). 2015-2017. Organismo responsable: MINECO.

3. Proyectos de **ámbito autonómico**:

   - **2014 SGR 1055**: *"Ajuts per donar suport a les activitats dels grups de recerca. Grups de recerca reconeguts (SGR 2014-2016)"*. Investigadora principal: Dra. Montserrat Aguadé. (Universitat de Barcelona). 2014-2016. Organismo responsable: Generalitat de Catalunya.
   - **2017 SGR 1287**: *"Ajuts per donar suport a les activitats dels grups de recerca. Grups de recerca reconeguts (SGR 2017-2019)"*. Investigadora principal: Dra. Montserrat Aguadé. (Universitat de Barcelona). 2017-2019. Organismo responsable: Generalitat de Catalunya.